Universiteit Gent
Faculteit Wetenschappen
Vakgroep Geografie

Sensing and making sense of crowd dynamics using *Bluetooth tracking*

An application-oriented approach

Dissertation

Meten en begrijpen van groepsverplaatsingen door het gebruik van *Bluetooth tracking*

Een toepassingsgerichte benadering

Proefschrift

Mathias Versichele

Proefschrift aangeboden tot het behalen van de graad van doctor in de wetenschappen: geografie

iwt
agentschap voor Innovatie
door Wetenschap en Technologie

Promotoren:
Prof. Dr. Nico Van de Weghe
Universiteit Gent
Prof. Dr. ir. Ingrid Moerman
Universiteit Gent

Leden van de leescommissie:
Prof. Dr. Gennady Andrienko
Fraunhofer Institute for Intelligent Analysis and Information Systems
City University London
Prof. Dr. ir. Stefan van der Spek
Technische Universiteit Delft
Prof. Dr. ir. Serge Hoogendoorn
Technische Universiteit Delft
Dr. Tijs Neutens
Universiteit Gent

Overige leden van de examencommissie:
Prof. Dr. Philippe De Maeyer
Universiteit Gent
Prof. Dr. Frank Witlox
Universiteit Gent
Prof. Dr. ir. Sidharta Gautama
Universiteit Gent
Prof. Dr. Ben Derudder
Universiteit Gent

"You see, wire telegraph is a kind of a very, very long cat. You pull his tail in New York and his head is meowing in Los Angeles. Do you understand this? And radio operates exactly the same way: you send signals here, they receive them there. The only difference is that there is no cat."
- Albert Einstein

"Men are like Bluetooth, they're connected to you when you're nearby, but search for other devices when you're far away. Women are like WiFi, they see available devices but connect to the strongest one."
- Anonymous

Table of Contents

List of Figures

List of Tables

Preface

Looking forward to roam across South-America with my girlfriend, a fresh new Master's degree under my belt, but seriously doubting whether my life at Ghent University had come to an end. The time is July 2009, and I feel like I am at an important crossroad in my life. Do I follow the 'normal' procedure and look for a job, or do I take a leap into the unknown by responding to the repeated appeals of Nico and apply for a PhD grant in order to do research on what started out as a 'silly' idea one year before ? Flash forward to December 2013, and I am writing the preface of this dissertation in my recently purchased house, looking at my wife feeding my brand new son. Looking back, the last four years of my life have felt like a rollercoaster. Yet I don't regret a single thing!

Doing research on 'Bluetooth tracking' turned out be a one-of-a-kind experience, both for the good and the bad. Where the novelty of the approach and the lack of an academic foundation (or colleagues for that matter) often made me feel like the odd one out in academic circles, it also resulted in a large degree of freedom and a feeling of satisfaction of being at the forefront of developments. In fact, it was the novelty of the subject and its inherently technological character (which would certainly involve programming skills) which had me hooked from the onset. When the approach was first applied at the Rock Werchter Festival in 2009 (Master's thesis of Bram Van Londersele), the sudden media attention which ensued pointed out that the concept not only fascinated me (though it must be stated that some of this interest was also due to a fear of privacy infringements). Ever since, our line of research has revolved around a number of tracking projects. As time progressed and our technological capacities and experience grew, the motivation of these projects evolved from the mere collection of data to the deployment of a complete platform capable of providing real-time insights on crowds. As a result, I have had the privilege to work together with many people (inside, but largely outside the academic world) and get a good feeling of the societal aspects of Bluetooth tracking.

As in any other PhD, this dissertation would never have been written without the help of a number of individuals and institutions. First of all, I would like to gratefully acknowledge the Agency for Innovation by Science and Technology in Flanders (IWT) for believing in my research proposal and funding my work. Hopefully, they can play an important role in the future valorization of our line of research. Indirectly, I would also like to mention the Research Foundation-Flanders (Belgium) (FWO) for funding some of my co-authors.

Besides funding, I also owe gratitude to a multitude of persons which I have had the pleasure to collaborate with over the last four years. First, I am deeply obliged to my supervisor Nico Van de Weghe. In fact, without his enthusiasm when explaining the research proposal I probably wouldn't be where I am right now. Of his many qualities, I have especially appreciated his down-to-earth mentality, his personal involvement in my research, and the speed and flexibility with which he has counseled me. My other supervisor, Ingrid Moerman, I would like to thank for her guidance in my research proposal and the technical support (both personally and through her collaborators) in the early stages of our research. A special mention should also go to Tijs Neutens. Over the last four years, he was often the first person I consulted whenever I had questions. Without his academic excellence (including linguistic aspects) this dissertation would essentially not be the same.

As this PhD represents a collection of academic articles, I would also like to gratefully acknowledge my co-authors: Nico Van de Weghe, Ingrid Moerman, Tijs Neutens, Matthias Delafontaine, Stephanie Goudeseune, Frederik Van Bossche, Manuel Claeys Bouùàert and Liesbeth De Grootte.

I would also like to thank my colleagues from 'the S8' over the last four years, both those that I shared an office with — Roel, Manuel, Pengdong, Hossein, Yi, Matthias, Frederik — and the rest (not in order of importance): Soetkin, Ann, Rasha, Tijs, Wim, Bart Dewulf, Bart De Wit, Marijke, Ruben, Katrien, Timo, Coen, Berdien, Pepijn, Joke, Lieselot, Kristien, Sabine, Karine, Eline, Luc, Valerie, Lander, Koen, Ellen, Annelies, Bram, Els, Ilke, Kobe, Veronique, Elien, Nathalie, Anneleen. Helga I would like to acknowledge for her countless administrative skills, and the occasional small-talk in her office.

Over the last four years, I have also had the pleasure to accompany ten Master's students on the way to successfully completing their thesis: Bram Van Londersele, Tim Claeys, Edward De Mûelenaere, Sara De Roeck, Frederik Van Bossche, Tom Baeyens, Roel Huybrechts, Stefaan De Vos, Stephanie Goudeseune and Liesbeth De Grootte. A special thank you go to Stephanie for her work during the 'Tour of Flanders 2011' and to Liesbeth for her field work in May 2012. Both these datasets resulted in a chapter of this dissertation. To all of them: the best of luck in whatever you intend on doing in the future.

As previously mentioned, a large number of people outside of the academic confines of Ghent University have directly or indirectly contributed to this dissertation. While it is impossible to name all of them, I want to give credit where it is due. For the permission and support during the experiments at the 'Ghent Festivities' 2010–2013, I would like to thank Lieven Decaluwe, Christophe Peeters, mayor Daniel Termont, Wim Beelaert, Jan Schiettekatte, Heidi Rogiest, Jeroen De Schuyteneer, Nikolai Abramovich, and local square organizers Ivan Saerens, Katleen Haentjens, Joris De Wildeman, Henk Dedeurwaerder ('Onder de draak), Guido De Leeuw ('Trefpunt'), Ilse Everaert ('Uitbureau Gent'), Jo Bonte ('Polé polé'), Nancy De Bruyne ('Happy days'), Eli and Tine De Ryck ('Koninklijke Dekenij Korenmarkt'), Gerald Claes ('Charlatan'), and Nico Cremers ('GRC'). A special mention goes to Christiaan De Pauw, leftenant at the fire department, whom we consider as our number one fan. For the 'Ghent Light Festival' in 2012, we would like to thank Kaat Heirbrant, Pieter François and Serge Platel. For the 'Tour of Flanders' project in 2011, we would like to thank Wim Van Herreweghe and Carlos De Vos. Geert Laureys and Yo Van Saet ('Live Nation') are gratefully acknowledged for their permission to perform experiments during the 'Rock Werchter' and 'I Love Techno' events. I would also like to thank 'Artexis', and in particular Luc Paeme and Nancy Lippens, for their permission and financial support in testing our methodology on the Horeca Expo events. Even though I was not present because I was to busy marrying my wife, I would also like to thank some colleagues from Hasselt University for their work at the Pukkelpop festival: Bram Bonné, Arno Barzan, Peter Quax and Wim Lamotte. Egbert Jaspers, Rob Wijnhoven and Kris van Rens ('ViNotion') are gratefully acknowledged for the use of their automatic people counting system during the 'Ghent Festivities 2013'. Tom De Jaeger from 'Tersec' is gratefully acknowledged for the use of his camera system on top of the Belfry during the same event, as are the brave firemen who installed it. Finally, we also appreciate all persons which allowed us to install a Bluetooth scanner in their premises during events, but for the sake of the environment — and the rainforest in particular — I will not list all of them explicitly.

A very important acknowledgement is devoted to Roel Huybrechts. Without his technical background and assistance, I and all others which have worked or are working on the same subject would not have made it this far. I look forward to teaming up with him to transform our product from 'nice-to-have' to 'need-to-have'!

Finally, I would like to thank my family for their continuing support and love. 'Papa', 'mama', Emanuelle, Vincent: thank you for just being who you are. Camiel and Alixe: thank you for bringing joy and happiness to the family as only children can accomplish. Karen, my dear wife and love of my life, thank you for ending up in Belgium and turning my world upside down! Most of all, thank you for giving me the most beautiful gift imaginable last September: a healthy son. Lucas: you might not imagine it yet; but as one day you might get your hands on this dusty book of your 'old man', know that the importance of a PhD means nothing in comparison to your well-being.

Ghent, November 2013
Mathias Versichele

1 Introduction

1.1 Background and motivation

1.1.1 Crowd dynamics and their relevance

Over the years, crowds and their dynamic nature have been studied in a wide area of research fields. Pedestrian behavior at bottlenecks in the built environment, for example, has important implications for a crowd's safety in case of evacuations (Daamen and Hoogendoorn, 2011; Helbing et al., 2000). Various incidents at mass events have unfortunately shown that mismanagement of such crowds, panic situations, riots or meteorological events can lead to disasters with significant casualties (Dickie, 1995; Helbing and Mukerji, 2012; Zhen et al., 2008). When interpreted more loosely as a group of of individuals with similar intentions, the relevance of crowd movements becomes more diverse. In retail environments, for example, crowds represent potential buyers and the routes they take have profound impacts from a micro-economic perspective. As a result, several efforts have already been reported to measure, analyze and model shopping behavior, both within stores (Hui et al., 2009a,b; Larson et al., 2005) and on a larger scale such as a city (Borgers and Timmermans, 1986; Kemperman et al., 2009). Tourists represent another example of a group (or 'crowd') with similar goals, and a rich vain of research on their spatiotemporal behavior has grown rapidly due to the emergence of positioning technologies such as GPS (Shoval and Isaacson, 2009).

1.1.2 The intangible nature of crowds

Crowd behavior is often simulated, usually either by agent-based models (Batty, 2005; Batty et al., 2003; Helbing et al., 2000, 2007; Helbing and Molnar, 1995; Hoogendoorn and Bovy, 2004; Pelechano et al., 2007) or cellular automata (Blue and Adler, 2001; Burstedde et al., 2001; Dijkstra et al., 2001; Kirchner and Schadschneider, 2002). Regardless of the context and their interpretation, crowds are complex phenomena and any attempt at modeling them ultimately needs to be validated through empirical movement data. Measuring movements of crowds and more specifically of the individuals constituting them is not a trivial task, however, for several reasons. First, crowd movements are usually associated with relatively short pedestrian trips which, in contrast with larger distance transport modes such as air or rail transport, cannot be traced or reconstructed from statistical data sources. Second, small-scale movements are only meaningful if their level of detail (i.e. frequency and accuracy of location registrations) corresponds to the level of detail of the environment they take place in. Third, crowds are ephemeral entities — in contrast to communities — which negatively influences the feasibility of contacting individuals composing them to directly participate in studies monitoring their movement. Finally, crowds are per definition composed of a large number of people and can only be aptly interpreted by studying a sufficiently large and representative subset of these individuals which in turn further complicates any direct involvement of studied individuals.

These characteristics function as constraints for any potential methodology aiming to capture crowd movements. Conventional methodologies such as the shadowing of individuals (Millonig and Gartner, 2011; Quinlan, 2008), direct interviews or the use of trip diaries (Axhausen et al., 2002; Thornton et al., 1997) are ill-suited for studying crowds as they are too labor-intensive to scale up from small groups to large crowds. In contrast, observation through cameras in public spaces does not require the direct cooperation of tracked individuals. In recent years, substantial progress has been made in *inter alia* the automatic detection of moving objects (Rabaud and Belongie, 2006; Viola et al., 2005), crowd behavior recognition (Saxena et al., 2008), and crowd density estimation (Marana et al., 1998). The use of video data to track individual movements within crowds remains a highly challenging task, however, due to dense packing and constant interactions among individuals inside a crowd, limited viewing angles and changes of illumination and weather conditions. More importantly, larger areas need to be covered by several cameras, and the reconstruction of individual movements across multiple camera views under realistic conditions is an unsolved problem to date. Hence, current applications of video surveillance have achieved to record the spatiotemporal paths of only few objects in limited spatial environments (Dee and Velastin, 2007). As a result, video technology is capable of capturing small-scale movements within one camera view (Daamen and Hoogendoorn, 2003; Johansson et al., 2008) but unable to study crowds moving over larger areas.

1.1.3 Measuring through 'proxies'

In contrast with the direct observation of individuals or through video analysis, objects that are somehow associated with or carried around by persons are often easier to locate through space and time. A surprising variety of these 'proxies' for human movement have already been studied, including one-dollar bank notes (Brockmann et al., 2006), trackable items or 'travel bugs' used in geocaching (Brockmann and Theis, 2008), and public transit smart cards (Pelletier et al., 2011). Global navigation satellite systems provide a better-known alternative. After GPS (global positioning system) was announced fully operational for civilian use in 1995, it took until the turn of the century for the distribution of GPS logging devices to become more or less mainstream in empirical research (Draijer et al., 2000; Murakami and Wagner, 1999). Over the years, the number of implementations grew quickly as logging devices became less expensive and less dependent on active participation of the studied individual during the tracking period (Andrienko et al., 2013; Bohte and Maat, 2009; Van der Spek et al., 2009; Van Schaick and Van der Spek, 2008; Vazquez-Prokopec et al., 2009).

In the end, however, it is the mobile phone that is starting to truly revolutionize empirical research into human mobility as they are equipped with a wide variety of location-aware technologies (including GPS), and are usually always in close contact with their owners. With around 6 mobile phone subscriptions for every 5 inhabitants of developed countries, and around 86 per 100 inhabitants worldwide in 2011 (International Telecommunication Union, 2012), the potential of using mobile phones as proxies for measuring the movement of large populations is clearly unparalleled. As a result, several methodologies making use of mobile phones have gained considerable attention. Roughly, these can be subdivided into participatory (needing direct collaboration of the mobile phone owner) handset-based solutions and non-participatory network-based solutions. Handset-based solutions, where the position of the mobile device is determined on the device itself (usually through GPS), typically involve smartphone applications gathering GPS (and other) data and backend servers collecting, processing and storing these data. Data gathered in this way include geo-tagged pictures (Girardin and Calabrese, 2008; Jankowski et al., 2010), shared locations from platforms such as Foursquare (Cheng et al., 2011), shared bike trips (Charlton and Schwartz, 2010) and shared locations during a mass event (Wirz et al., 2013). Alternatively, network-based methodologies can be used to derive the location of mobile phones without any participation of their owners by analyzing datasets of mobile phone operators and deriving which cell towers phones were connected to when performing a certain activity such as making a call or sending a text message. Ever since the first documented use in Milan in 2006 (Ratti et al., 2006), an increasing number of studies have employed this methodology for both applied research into tourist (Ahas et al., 2007a, 2008) and commuter dynamics (Ahas et al., 2010) as well as more fundamental research into human mobility patterns (Candia et al., 2008; González et al., 2008; Song et al., 2010), geographic borders (Blondel et al., 2010) and privacy implications (de Montjoye et al., 2013).

At first sight, non-participatory network-based mobile phone tracking represents a nearperfect match for tracking crowds. Locations of phones, however, are usually limited to the locations of cell towers the phones were connected to during the registration. A study in Estonia, for example, demonstrated that around 50% of measurements were correct to within only 400 meters in urban areas and only 2600 meters in rural areas (Ahas et al., 2007b). Additionally, all reported datasets in the literature only provided locations of phones when a call or message activity was made. As such, the methodology fails to provide the level of detail necessary for capturing smaller-scale movement patterns — both in the spatial as well as the temporal dimension. The spatiotemporal coverage and quality of handset-based GPS datasets, on the other hand, is highly sensitive to the degree of cooperation of phone owners.

1.1.4 Bluetooth tracking as a potential alternative

The overview in the previous section indicates that there is no established methodology for the non-participatory measurement of crowd movements with a spatial granularity finer than that of a cell tower network. Recently, alternative approaches using short-range wireless technologies have been proposed. Although there are a number of candidate technologies such as WiFi (Bonné et al., 2013) and RFID (Kanda et al., 2007), Bluetooth has gained particular attention over the last few years. Bluetooth, originally developed in Sweden by Ericsson in 1994, is a wireless communication technology designed for ad-hoc exchange of data and information between mobile devices. Ever since its conception, it has quickly become a quasi ubiquitous mobile technology available on an estimated two billion devices worldwide (Bluetooth Special Interest Group, 2012). Prior to being able to set up a connection between two devices, a 'master' device first needs to inquire for the presence of a nearby 'slave' device. When the Bluetooth functionality of the slave device is turned on and set to 'discoverable', it broadcasts its MAC address to the inquiring master device. This 48-bit identifier uniquely identifies the Bluetooth module of the device. Although the technology was originally only conceived as a means of replacing wired connections necessary for file exchange or synchronization between devices, the idea began materializing over the years that the 'proximity' (Bensky, 2007) sensing of mobile devices could also have other applications.

In general, two such applications which gained considerable scientific attention are interaction modeling and positioning. *Interaction modeling* approaches deduce the potential for interaction between individuals from the physical proximity of the individuals' devices. This has caused a diverse range of research avenues into reconstructing social networks (Eagle and Pentland, 2005; Nicolai et al., 2006), inferring shared interests (Terry et al., 2002), investigating 'familiar strangers' (Paulos and Goodman, 2004), developing social mobile services (Rudström et al., 2004), and even modeling political opinions (Madan et al., 2011). Bluetooth interaction modeling is also relevant to more technically oriented research questions concerned with ad-hoc forwarding algorithms (Barzan et al., 2013; Hui et al., 2005) or delay-tolerant networks (Natarajan et al., 2007). Besides interaction modeling, another application is the *positioning* of mobile devices. Surveying the literature, it becomes clear that approaches to this end apply two alternative approaches. Since the turn of the last century, Bluetooth has been intensively studied for indoor location tracking purposes as an alternative for GPS technology (Feldmann et al., 2003; Hallberg et al., 2003; Madhavapeddy and Tse, 2005; Rodriguez et al., 2005). These approaches are mainly handset-based where devices calculate their own position with reference to a number of static base stations, usually by not only detecting proximity but also registering the RSSI (received signal strength intensity) of detections (Pei et al., 2010). As this signal strength is negatively correlated with the distance between the inquiring and detected device, it can theoretically be used to estimate distances (Hossain and Soh, 2007) and hence also locations by multilateration (Bensky, 2007).

A second approach involves deploying Bluetooth sensors over a study area and reconstructing movements by matching the MAC addresses of detected devices with the locations of the sensors they were detected by. This network-based and non-participatory approach started materializing after a first documented trial in 2006 in Bath, UK (O'Neill et al., 2006). Since then, a growing number of experimental use-cases have been documented. Particular attention has been devoted to the use of Bluetooth technology for travel time measurements of motorized traffic as it represents a simplified approach in comparison to either number plate recognition or GPS floating car data (Haghani et al., 2009; Hamedi et al., 2010; Malinovskiy et al., 2011; Martchouk et al., 2010; Wasson et al., 2008). Pedestrian mobility has also been investigated. Examples include transit time measurements in airport security checkpoints (Bullock et al., 2010), travel and dwelling time calculations in an urban context (Malinovskiy et al., 2012), and the automatic registration of public transport users (Kostakos et al., 2013; Weinzerl and Hagemann, 2007). The 'Cityware' project in the United Kingdom used static Bluetooth sensors to capture mobility traces, and coupled these data with user's online social data (Kostakos and O'Neill, 2008). The majority of deployments use static sensors, but proof-of-concept demonstrations using smartphones as mobile sensors have already been documented (Malinovskiy and Wang, 2012; Morrison et al., 2009; Stopczynski et al., 2013). Because of the diversity in application domains, there does not seem to be a consensus on a common denominator for this non-participatory detection of Bluetooth devices. Following Van Londersele et al. (2009) and Leitinger et al. (2010), the methodology will be referred to as 'Bluetooth tracking' in the remainder of this dissertation.

Due to the non-participatory nature of the approach (no participation is necessary of the tracked individuals) and the flexibility in the locations where sensors are deployed, Bluetooth tracking seems particularly promising for collecting mobility traces at mass events. Not surprisingly, a number of experiments at mass events have already been described in the last few years. Leitinger et al. (2010) presented a first concise experiment during the Donauinselfest 2009 in Austria. Stange et al. (2011) described the analytical workflow of using Bluetooth to monitor human mobility during a Formula 1 race. Andrienko et al. (2012) used Bluetooth tracking data to highlight the specific characteristics of spatiotemporally sparse or 'episodic' movement data, and how visual analytics can aid in correctly interpreting these data. Larsen et al. (2013) investigated musical preferences of visitors of the Roskilde festival in Denmark by registering them with Bluetooth sensors near the different stages. Stopczynski et al. (2013) demonstrated a proof-of-concept of using smartphones as mobile Bluetooth sensors at the same festival.

1.2 Objectives and approach of the study

Over the last few years, Bluetooth tracking has emerged as a promising approach in capturing crowd movements in a non-participatory manner and on a level of detail which cannot be delivered by records of mobile phone operators. While the number of use-cases at mass events is gradually increasing, it remains unclear what the ultimate potential of the methodology is. By building on the limited body of research on Bluetooth tracking to date, this dissertation has the following objectives:

[i] to comprehensively illustrate and document the benefits and issues of Bluetooth tracking at mass events;

[ii] to explore the potential for applications outside the scope of mass events; and

[iii] to investigate the process of analyzing Bluetooth tracking data and their specific characteristics.

The approach in doing so is application-oriented, where the contribution of each chapter in the dissertation is empirically illustrated on a different Bluetooth tracking dataset. This dissertation is composed of a collection of four academic articles (two of which are published, one accepted and forthcoming, and one currently under a review process). The general outline is depicted in figure 1.1 on the facing page. Together, these four chapters represent the most important research results over the last four years[1 2]. Chapter 6 on page 121 finally summarizes these research results, with a special emphasis on issues which require further research and privacy implications.

In parallel to the publication of these results, substantial efforts and time have also been devoted to developing a platform capable of gathering, centralizing, and analyzing (both offline and in real-time) Bluetooth tracking data on a large scale. One of these efforts has resulted in the *GISMO* toolkit for offline analysis of Bluetooth tracking data. Appendix A on page 141 provides both a background on the necessity for such a toolkit and an overview of its main capabilities by processing a dataset gathered at a music festival. The development

[1]Due to very recent developments in the field of Bluetooth tracking and other similar fields, the bibliographies of some of the chapters which were published earlier are somewhat outdated. For the most comprehensive and up-to-date overview, readers are referred to the bibliography of this introduction.

[2]Because all chapters are separately published (or in the process of being published) in different outlets, some repetition between them is unavoidable (in particular with reference to literature studies and the working principle of the Bluetooth tracking methodology).

of the *GISMO* toolkit addresses the third objective of this dissertation. As depicted by the dashed lines in figure 1.1, the toolkit served an important supportive role for this dissertation by aiding in the analysis of datasets gathered from static sensor networks in chapters 2, 4 and 5.

Four years and 20 Bluetooth tracking projects have not only resulted in this dissertation, but also in several oral presentations, a book chapter, reports and ten Master's Theses which I supervised. As a reference, appendix B on page 165 provides a summarized overview of these auxiliary accomplishments.

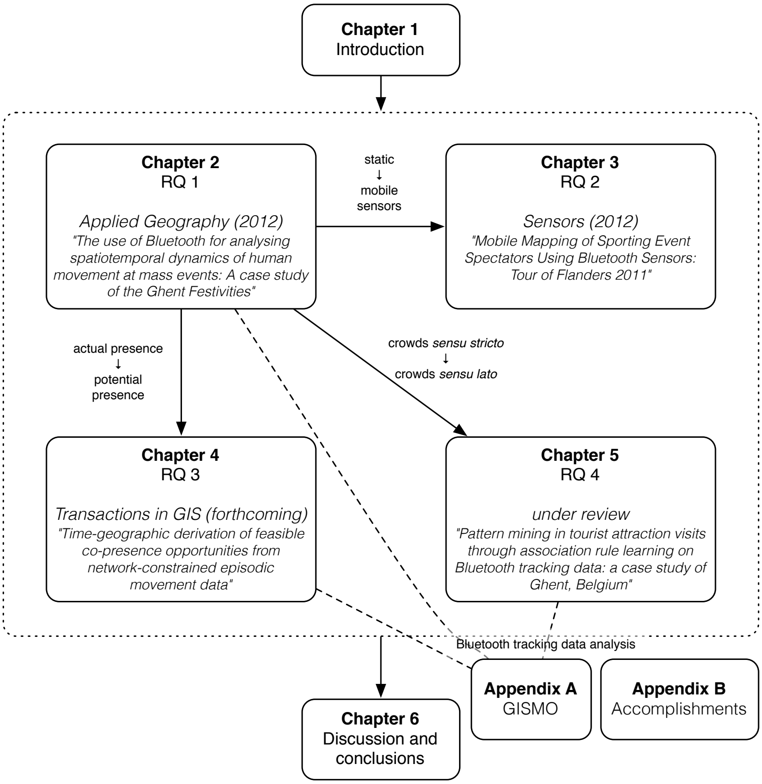


Figure 1.1: General outline of the thesis.

1.3 Research questions

The general aims of this dissertation have crystallized into four identifiable research questions, each of which is addressed in a separate chapter. In this section, we give a brief summary.

RQ 1: Which opportunities does Bluetooth tracking provide for studying spatiotemporal dynamics within crowds at mass events?

This research question addresses the first objective of this dissertation. Chapter 2 on page 19, published in *Applied Geography* (Versichele et al., 2012a), documents the use of Bluetooth tracking at the 'Ghent Festivities 2010' event in Belgium which attracted around 1.5 million visitors over ten days. In contrast to existing scientific contributions, the focus lies on the methodology itself and on the broad analytical potential in the context of mass events. The discussion also includes details on the deployed hardware (22 Bluetooth sensors) and software, which aims to make the approach more tangible to a general audience. Instead of performing one specific analysis, the chapter provides an overview of analytical possibilities by empirical examples on the dataset which contained more than 150,000 trajectories. These include estimations of crowd sizes at sensor locations, flow sizes in between these locations, and the usage of public transport by the visitors. The chapter concludes by discussing both the advantages as well as the remaining issues of the Bluetooth tracking methodology.

RQ 2: Can Bluetooth technology be used to count and map complex crowds dispersed over large areas?

Current implementations of Bluetooth tracking usually studied crowds and their dynamics confined to certain event zones (Larsen et al., 2013; Leitinger et al., 2010; Stange et al., 2011; Versichele et al., 2012a). As the geographical extent of a crowd grows, however, the feasibility of studying it by a network of static sensors degrades. Examples of such events are sport events which take place in public space. This research question takes a step back from RQ 1 by not focusing on mobility or dynamics within crowds as such, but on counting (determining the crowd size) and mapping (determining its geographic extent) these complex crowds. Chapter 3 on page 45, published in *Sensors* (Versichele et al., 2012b), starts addressing the research question by giving an overview of the state-of-the-art in counting and mapping of crowds. As convential methodologies are problematic due to scaling or feasibility issues, we propose using Bluetooth sensors installed on a mobile platform. A case study is described where spectators of a cycling race ('Tour of Flanders 2011') are mapped by these Bluetooth sensors moving along the race track. To the best of our knowledge, this is the first reported application of the mobile mapping paradigm with Bluetooth technology to count and map a geographically spread crowd.

RQ 3: How can a crowd's movement within sensor locations be modeled?

A key characteristic of Bluetooth tracking data — especially in comparison to GPS data — is its spatiotemporal sparseness. Because sensors are only present at certain locations in a study area, trajectories are often made up of long time gaps where the location of an individual is unknown due to his/her device being out of reach of all sensors. Chapter 4 on page 67, accepted for publication in *Transactions in GIS*, explores the possibility of modelling an entity's potential path in between two consecutive detections by different sensors. Using the time-geographical framework, it develops a model able to derive potential co-presence of individuals within a crowd. The model is applied on a dataset from the 'Ghent Light Festival 2012', and used to generate a time-series view of the potential spread of the crowd within sensor locations. By focusing on the spatiotemporally sparse or episodic (Andrienko et al., 2012) nature of Bluetooth tracking data, this research question adresses the third objective of this dissertation.

RQ 4: What is the value of Bluetooth tracking outside of the context of mass events?

This rather general research question is addressed by chapter 5 on page 89, currently under review. The chapter proposes the use of Bluetooth tracking in a tourism context and thus formulates an answer to the second objective of the dissertation. In contrast to the previous chapters it lets go of the strict definition of crowds, as already discussed in the introduction. An experiment is described where visits to tourist attractions over fifteen days in Ghent (Belgium) were tracked by Bluetooth sensors. The chapter first elaborately discusses the need for a filtering procedure to extract actual visits from the tracking data, and how the deployment of sensors in hotels and the tourist inquiry desk can provide additional context. The dataset of visits is subsequently mined by an association rule learning scheme, able to detect which attractions are more often combined than usual by certain visitor segments.

References

Ahas, R., Aasa, A., Mark, U., Pae, T., and Kull, A. (2007a). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898–910.

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies*, 18(1):45–54.

Ahas, R., Laineste, J., Aasa, A., and Mark, U. (2007b). The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In Gartner, G., Cartwright, W., and Peterson, M. P., editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 445–460. Springer, Berlin.

Andrienko, N., Andrienko, G., Barrett, L., Dostie, M., and Henzi, P. (2013). Space transformation for understanding group movement. *IEEE transactions on visualization and computer graphics*, 19(12):2169–78.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., and Hecker, D. (2012). Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. *KI - Künstliche Intelligenz*, 26(3):241–251.

Axhausen, K., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.

Barzan, A., Bonné, B., Quax, P., Lamotte, W., Versichele, M., and Van de Weghe, N. (2013). A comparative simulation of opportunistic routing protocols using realistic mobility data obtained from mass events. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Madrid. IEEE.

Batty, M. (2005). Agents, cells, and cities: new representational models for simulating multiscale urban dynamics. *Environment and Planning A*, 37(8):1373–1394.

Batty, M., Desyllas, J., and Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, 17(7):673–697.

Bensky, A. (2007). *Wireless positioning technologies and applications.* Artech House, Boston, London.

Blondel, V., Krings, G., and Thomas, I. (2010). Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42:1–12.

Blue, V. J. and Adler, J. L. (2001). Cellular automata microsimulation for modeling bi-directional pedestrian walkways. *Transportation Research Part B: Methodological*, 35(3):293–312.

Bluetooth Special Interest Group (2012). 2012 Annual Report. Technical report.

Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297.

Bonné, B., Barzan, A., Quax, P., and Lamotte, W. (2013). WiFiPi: Involuntary tracking of visitors at mass events. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Madrid.

Borgers, A. and Timmermans, H. (1986). A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas. *Geographical analysis*, 18(2):115–128.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Brockmann, D. and Theis, F. (2008). Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns. *Pervasive Computing, IEEE*, 7(4):28–35.

Bullock, D. M., Haseman, R., Wasson, J., and Spitler, R. (2010). Anonymous Bluetooth Probes for Measuring Airport Security Screening Passage Time: The Indianapolis Pilot Deployment. *Transportation Research Board*, pages 1–16.

Burstedde, C., Klauck, K., Schadschneider, A., and Zittartz, J. (2001). Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications*, 295(3-4):507–525.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Charlton, B. and Schwartz, M. (2010). CycleTracks – a Bicycle Route Choice Data Collection Application for GPS-Enabled Smart Phones. In *3rd International Conference on Innovations in Travel Modeling (ITM2010) - Transportation Research Board (TRB)*, number 1, page 6, Tempe, AZ.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 81–88, Barcelona.

Daamen, W. and Hoogendoorn, S. (2003). Experimental Research of Pedestrian Walking Behavior. *Transportation Research Record*, 1828(1):20–30.

Daamen, W. and Hoogendoorn, S. (2011). Emergency Door Capacity: Influence of Population Composition and Stress Level W. In Peacock, R. D., Kuligowski, E. D., and Averill, J. D., editors, *Pedestrian and Evacuation Dynamics*, pages 15–24. Springer US, Boston, MA.

de Montjoye, Y.-A., Hidalgo, C. a., Verleysen, M., and Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1376):1–5.

Dee, H. M. and Velastin, S. A. (2007). How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, 19(5-6):329–343.

Dickie, J. (1995). Major crowd catastrophes. *Safety Science*, 18(4):309–320.

Dijkstra, J., Timmermans, H., and Jessurun, A. (2001). A Multi-Agent Cellular Automata System for Visualising Simulated Pedestrian Activity. In Bandini, S. and Worsch, T., editors, *Theory and Practical Issues on Cellular Automata - Proceedings of the Fourth International Conference on Cellular Automata for Research and Industry*, pages 29–36, Karlsruhe. Springer.

Draijer, G., Kalfs, N., and Perdok, J. (2000). Global Positioning System as Data Collection Method for Travel Research. *Transportation Research Record*, 1719(1):147–153.

Eagle, N. and Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Feldmann, S., Kyamakya, K., Zapater, A., and Lue, Z. (2003). An indoor Bluetooth-based positioning system: concept, implementation and experimental evaluation. In *International Conference on Wireless Networks (ICWN)*, pages 109–113, Las Vegas, NV.

Girardin, F. and Calabrese, F. (2008). Digital Footprinting: Uncovering Tourists with User-Generated Content. *Pervasive Computing*, 7(4):36–43.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., and Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160:60–68.

Hallberg, J., Nilsson, M., and Synnes, K. (2003). Positioning with Bluetooth. In *10th International Conference on Telecommunications (ICT 2003)*, pages 954–958, Papeete. IEEE.

Hamedi, M., Fish, R., and Haghani, A. (2010). Freeway Dynamic Message Sign Evaluation Using Bluetooth Sensors: A Case Study. In *17th ITS World Congress*, pages 1–14, Busan.

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490.

Helbing, D., Johansson, A., and Al-Abideen, H. (2007). Dynamics of crowd disasters: An empirical study. *Physical Review E*, 75(4).

Helbing, D. and Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282–4286.

Helbing, D. and Mukerji, P. (2012). Crowd disasters as systemic failures: analysis of the Love Parade disaster. *EPJ Data Science*, 1(1):7.

Hoogendoorn, S. and Bovy, P. (2004). Pedestrian route-choice and activity scheduling theory and models. *Transportation Research Part B*, 38(2):169–190.

Hossain, A. and Soh, W. (2007). A comprehensive study of Bluetooth signal parameters for localization. In *18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, Athens, Greece.

Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J., and Diot, C. (2005). Pocket switched networks and human mobility in conference environments. In *Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking (WDTN '05)*, pages 244–251, New York, NY. ACM Press.

Hui, S. K., Bradlow, E. T., and Fader, P. S. (2009a). Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior. *Journal of Consumer Research*, 36(3):478–493.

Hui, S. K., Fader, P. S., and Bradlow, E. T. (2009b). The Traveling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP Optimality. *Marketing Science*, 28(3):566–572.

International Telecommunication Union (2012). Measuring the Information Society. Technical report, Geneva.

Jankowski, P., Andrienko, N., Andrienko, G., and Kisilevich, S. (2010). Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transactions in GIS*, 14(6):833–852.

Johansson, A., Helbing, D., Al-Abideen, H. Z., and Al-bosta, S. (2008). From Crowd Dynamics to Crowd Safety: A Video-Based Analysis. *Advances in Complex Systems*, 11(4):497–527.

Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H., and Hagita, N. (2007). Analysis of People Trajectories with Ubiquitous Sensors in a Science Museum. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 4846–4853, Rome. IEEE.

Kemperman, A., Borgers, A., and Timmermans, H. (2009). Tourist shopping behavior in a historic downtown area. *Tourism Management*, 30(2):208–218.

Kirchner, A. and Schadschneider, A. (2002). Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. *Physica A: Statistical Mechanics and its Applications*, 312(1-2):260–276.

Kostakos, V., Camacho, T., and Mantero, C. (2013). Towards proximity-based passenger sensing on public transport buses. *Personal and Ubiquitous Computing*, 17(8):1807–1816.

Kostakos, V. and O'Neill, E. (2008). Capturing and visualising Bluetooth encounters. In *Adjunct proceedings of the Conference on Human Factors in Computing Systems (CHI 2008)*, Florence.

Larsen, J. E., Sapiezynski, P., Stopczynski, A., Moerup, M., and Theodorsen, R. (2013). Crowds, Bluetooth, and Rock-n-Roll. Understanding Music Festival Participant Behavior.

Larson, J. S., Bradlow, E. T., and Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing*, 22(4):395–414.

Leitinger, S., Gröchenig, S., Pavelka, S., and Wimmer, M. (2010). Erfassung von Personenströmen mit der Bluetooth-Tracking- Technologie. In *Angewandte Geoinformatik 2010*, pages 220–225, Salzburg, Austria.

Madan, A., Farrahi, K., Gatica-Perez, D., and Pentland, A. S. (2011). Pervasive Sensing to Model Political Opinions in Face-to-Face Networks. In Lyons, K., Hightower, J., and Huang, E. M., editors, *Pervasive Computing*, volume 6696 of *Lecture Notes in Computer Science*, pages 214–231. Springer, Berlin, Heidelberg.

Madhavapeddy, A. and Tse, A. (2005). A study of bluetooth propagation using accurate indoor location mapping. In *UbiComp 2005: Ubiquitous Computing*, volume 3660 of *Lecture Notes in Computer Science*, pages 105–122. Springer, Berlin, Heidelberg.

Malinovskiy, Y., Lee, U.-K., Wu, Y.-J., and Wang, Y. (2011). Investigation of Bluetooth-Based Travel Time Estimation Error on a Short Corridor. In *Transportation Research Board 90th Annual Meeting*, volume 2250.

Malinovskiy, Y., Saunier, N., and Wang, Y. (2012). Pedestrian Travel Analysis Using Static Bluetooth Sensors. In *Transportation Research Board 91st Annual Meeting*, volume 250.

Malinovskiy, Y. and Wang, Y. (2012). Pedestrian Travel Pattern Discovery Using Mobile Bluetooth Sensors. In *Transportation Research Board 91st Annual Meeting*, volume 250.

Marana, A., Velastin, S., Costa, L., and Lotufo, R. (1998). Automatic estimation of crowd density using texture. *Safety Science*, 28(3):165–175.

Martchouk, M., Mannering, F., and Bullock, D. (2010). Analysis of Freeway Travel Time Variability Using Bluetooth Detection. *Journal of Transportation Engineering*, 137(10):697–704.

Millonig, A. and Gartner, G. (2011). Identifying motion and interest patterns of shoppers for developing personalised wayfinding tools. *Journal of Location Based Services*, 5(1):3–21.

Morrison, A., Bell, M., and Chalmers, M. (2009). Visualisation of spectator activity at stadium events. In *13th International Conference on Information Visualisation*, pages 219–226, Barcelona.

Murakami, E. and Wagner, D. (1999). Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies*, 7(2-3):149–165.

Natarajan, A., Motani, M., and Srinivasan, V. (2007). Understanding Urban Interactions from Bluetooth Phone Contact Traces. In *Passive and Active Network Measurement - 8th Internatinoal Conference, PAM 2007, Louvain-la-neuve, Belgium, April 5-6, 2007. Proceedings*, volume 4427 of *Lecture Notes in Computer Science*, pages 115–124. Springer, Berlin, Heidelberg.

Nicolai, T., Yoneki, E., Behrens, N., and Kenn, H. (2006). Exploring social context with the wireless rope. In Meersman, R., Tari, Z., and Herrero, P., editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, volume 4277 of *Lecture Notes in Computer Science*, pages 874–883. Springer.

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A., Penn, A., Fraser, D., and Jones, T. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *8th International Conference on Ubiquitous Computing (UBICOMP 2006)*, pages 315–332, Orange County, CA.

Paulos, E. and Goodman, E. (2004). The familiar stranger: anxiety, comfort, and play in public places. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, number 1, pages 223–230, Vienna, Austria.

Pei, L., Chen, R., and Liu, J. (2010). Using Inquiry-based Bluetooth RSSI Probability Distributions for Indoor Positioning. *Journal of Global Positioning Systems*, 9(2):122–130.

Pelechano, N., Allbeck, J., and Badler, N. (2007). Controlling Individual Agents in High-Density Crowd Simulation. In Metaxas, D. and Popovic, J., editors, *Proceedings of the 2007 ACM SIGGRAPH Symposium on Computer Animation*, pages 99–108.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Quinlan, E. (2008). Conspicuous Invisibility: Shadowing as a Data Collection Strategy. *Qualitative Inquiry*, 14(8):1480–1499.

Rabaud, V. and Belongie, S. (2006). Counting Crowded Moving Objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, volume 1, pages 705–711. IEEE.

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

Rodriguez, M., Pece, J., and Escudero, C. (2005). In-building location using bluetooth. In *Proceedings of the International Workshop on Wireless Ad Hoc Networks*, London.

Rudström, A. s., Svensson, M., Cöster, R., and Höök, K. (2004). Mobitip: Using bluetooth as a mediator of social context. In *Ubicomp 2004 Adjunct Proceedings*, page 2.

Saxena, S., Brémond, F., Thonnat, M., and Ma, R. (2008). Crowd behavior recognition for video surveillance. In *Advanced Concepts for Intelligent Vision Systems (ACIVS) Conference*, pages 970–981, Juan-les-Pins.

Shoval, N. and Isaacson, M. (2009). *Tourist mobility and advanced tracking technologies*, volume 19 of *Routledge Advances in Tourism*. Routledge, New York, London.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–21.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Stopczynski, A., Larsen, J. E., Lehmann, S., Dynowski, L., and Fuentes, M. (2013). Participatory Bluetooth Sensing: A Method for Acquiring Spatio-Temporal Data about Participant Mobility and Interactions at Large Scale Events. In *International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications 2013*, pages 242–247, San Diego, CA.

Terry, M., Mynatt, E., Ryall, K., and Leigh, D. (2002). Social Net: Using Patterns of Physical Proximity Over Time to Infer Shared Interests. In *Proceedings of Human Factors in Computing Systems (CHI'02)*, pages 816–817, Minneapolis.

Thornton, P., Williams, A., and Shaw, G. (1997). Revisiting time - space diaries: an exploratory case study of tourist behaviour in Cornwall, England. *Environment and Planning A*, 29(10):1847–1867.

Van der Spek, S., Van Schaick, J., De Bois, P., and De Haan, R. (2009). Sensing Human Activity: GPS Tracking. *Sensors*, 9(4):3033–3055.

Van Londersele, B., Delafontaine, M., and Van de Weghe, N. (2009). Bluetooth Tracking. *GIM International*, 23(11):23–25.

Van Schaick, J. and Van der Spek, S., editors (2008). *Urbanism on Track - Application of tracking technologies in urbanism*. IOS Press, Amsterdam.

Vazquez-Prokopec, G. M., Stoddard, S. T., Paz-Soldan, V., Morrison, A. C., Elder, J. P., Kochel, T. J., Scott, T. W., and Kitron, U. (2009). Usefulness of commercially available GPS dataloggers for tracking human movement and exposure to dengue virus. *International journal of health geographics*, 8:68.

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012a). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2):208–220.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., and Van de Weghe, N. (2012b). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10):14196–14213.

Viola, P., Jones, M., and Snow, D. (2005). Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision*, 63(2):153–161.

Wasson, J. S., Sturdevant, J. R., and Bullock, D. M. (2008). Real-Time Travel Time Estimates Using Media Access Control Address Matching. *ITE Journal (Institute of Transportation Engineers)*, 78(6):20–23.

Weinzerl, J. and Hagemann, W. (2007). Automatische Erfassung von Umsteigern per Bluetooth-Technologie. *Nahverkehrspraxis*, 3:18–19.

Wirz, M., Mitleton-kelly, E., Franke, T., Camilleri, V., Montebello, M., Roggen, D., Lukowicz, P., and Troster, G. (2013). Using Mobile Technology and a Participatory Sensing Approach for Crowd Monitoring and Management During Large-Scale Mass Gatherings. In Mitleton-Kelly, E., editor, *Co-evolution of Intelligent Socio-technical Systems*, Understanding Complex Systems, pages 61–77. Springer, Berlin, Heidelberg.

Zhen, W., Mao, L., and Yuan, Z. (2008). Analysis of trample disaster and a case study – Mihong bridge fatality in China in 2004. *Safety Science*, 46(8):1255–1270.

2

The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the ‘Ghent Festivities’

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N.

***Applied Geography*, 32(2):208–220**

Abstract *In this chapter, proximity-based Bluetooth tracking is postulated as an efficient and effective methodology for analysing the complex spatiotemporal dynamics of visitor movements at mass events. A case study of the ‘Ghent Festivities’ event (1.5 million visitors over ten days) is described in detail and preliminary results are shown to give an indication of the added value of the methodology for stakeholders of the event. By covering 22 locations in the study area with Bluetooth scanners, we were able to extract 152,487 trajectories generated by 80,828 detected visitors. Apart from generating clear statistics such as visitor counts, the share of returning visitors, and visitor flow maps, the analyses also reveal the complex nature of this event by hinting at the existence of several mutually different visitor profiles. We conclude by arguing why Bluetooth tracking offers significant advantages for tracking mass event visi-*

tors with respect to other and more prominent technologies, and outline some of its remaining deficiencies.

2.1 Introduction

In the last few years the representation and analysis of large volumes of trajectory information of objects moving through geographical space has become a major topic of interest in research domains such as geographical information science (Ahlqvist et al., 2010; Shaw et al., 2008), computer science (Bogorny et al., 2009; Orlando et al., 2007), visual analytics (Andrienko and Andrienko, 2007) and urbanism (Van Schaick and Van der Spek, 2008). This burgeoning academic interest has emerged as a result of the increased feasibility and affordability of collecting detailed data about spatiotemporal phenomena triggered by the widespread adoption of location-aware technologies. Past studies have focused on the movements of various kinds of objects including vehicles (Quiroga and Bullock, 1998), animals (Laube et al., 2007), bank notes (Brockmann et al., 2006) and typhoons (Terry and Feng, 2010), but the majority of research has been devoted to human movement in different contexts and at various scales. Some examples are the movements of athletes on a pitch (Laube et al., 2005), tourists on a regional (Ahas et al., 2008) and local scale (Kemperman et al., 2009; O'Connor et al., 2005; Shoval and Isaacson, 2007b), and customers in a supermarket (Hui et al., 2009). In these contexts, advanced tracking technologies complement more traditional qualitative methods, such as shadowing (Quinlan, 2008) and collecting travel diaries (Axhausen et al., 2002).

Within research on human behaviour, particular attention has been devoted to the collective behaviour of crowds at mass events such as street parades, festivals, public assemblies, sporting events, and exhibitions (Batty et al., 2003; Helbing et al., 2007; Zeitz et al., 2009). Tragic events such as the recent stampede during the 2010 edition of the 'Love Parade' in Duisburg show that it is vital to have accurate information on the spatiotemporal flow of visitors at mass events. However, the collection of quantitative movement data by the use of advanced tracking technologies, with GPS (global positioning system) as the main example, in these contexts raises serious issues concerning feasibility. Distributing and recollecting tracking units to a sufficient number of individuals within the crowd is indeed a labour-intensive and expensive process. Additionally, GPS usage in dense urban settings and inside buildings engenders problems due to signal distortion.

To date, video surveillance on the basis of closed circuit television (CCTV) has been the customary approach to capture human motion in crowded environments. Technological advances in the last decade have led to a large number of distinct research topics related to video surveillance including crowd density estimation, crowd behaviour monitoring, and face detection and recognition in crowds (Saxena et al., 2008). Despite substantial progress made in recent years, *however,* the use of video data to track individual movements within crowds remains a challenging task. First, dense packing and constant interactions among individuals make it difficult to unambiguously distinguish between individuals in a crowd. Second, limited viewing angles and changes of illumination and weather conditions highly complicate the visual recognition of individual spatial patterns from imagery. Finally, difficulties also arise with respect to the reconstruction of individual movements across multiple camera views, which is a necessity for larger study areas. Hence, current applications of video surveillance have achieved to record the spatiotemporal paths of only few objects in limited spatial environments (Dee and Velastin, 2007), inhibiting its use as a tracking technology in the context of mass events.

In response to these issues and given the ubiquity of Bluetooth-enabled devices such as mobile phones and personal digital assistants (PDA) carried around by their owners, Bluetooth technology has increasingly been suggested as a simple and low-cost alternative for the reconstruction of spatial behaviour (Bullock et al., 2010; Fatah and Mottram, 2007; Leitinger et al., 2010; Van Londersele et al., 2009; Versichele et al., 2010; Wasson et al., 2008). Designed as an open and wireless communication technology by Ericsson in 1994, Bluetooth has become a well-known and widely implemented standard for wireless exchange of data between devices. If devices are set to be 'discoverable', the movement of the devices — and by extension their users — can be reconstructed by means of a unique Media Access Control (MAC) address that gets broadcasted in the discovery process. Because this fixed MAC address cannot be linked to any personal information such as names or phone numbers (contrary to the 'friendly name' of the device), tracked individuals remain anonymous avoiding potential privacy infringements. Applications of Bluetooth tracking include the estimation of travel times on highway segments (Haghani et al., 2009; Wasson et al., 2008), public transport usage in Graz (Weinzerl and Hagemann, 2007), movement behaviour in a shopping mall (Millonig and Gartner, 2008), and functioning as an extension on social data gathered from Facebook in the 'Cityware' project (Kostakos and O'Neill, 2008).

Because Bluetooth allows for non-participatory, unannounced and simultaneous tracking of a large number of individuals, it is particularly useful to study visitor flows at mass events. Despite this potential, only pilot studies using Bluetooth tracking at mass events have been reported (Leitinger et al., 2010; Van Londersele et al., 2009). This chapter aims to significantly augment the current knowledge by reporting on a recent and comprehensive experiment using Bluetooth as a tracking technology. The experiment was carried out at the 'Ghent Festivities', one of the largest outdoor cultural events in Europe which lasts for ten days and attracts around 1.5 million visitors. This setting offers a challenging test bed in terms of crowd size, duration of the event and spatial extent of the study area. The aim of this case study was to explore the potential of Bluetooth tracking for studying the spatiotemporal dynamics of visitors at mass events by highlighting a selection of analytical possibilities with the gathered data and showing the corresponding results.

The remainder of the chapter is organised as follows. Section 2.2 on the next page gives a brief discussion of Bluetooth as a tracking technology. Section 2.3 on page 24 describes the background and experimental design of the case study, and in section 2.4 on page 27 we describe the preprocessing of the tracking data. In section 2.5 on page 28, we present the results of this study. Finally, we contextualize these results, argue why Bluetooth tracking has the potential to become a valuable methodology for studying the dynamics associated with mass events, and outline some of its remaining deficiencies in section 2.6 on page 37.

2.2 Bluetooth as a tracking technology

2.2.1 Working principle

Bluetooth is a short-range, low-power and open protocol for implementing Wireless Personal Area Networks (WPAN) between mobile devices. It operates on the Industrial, Scientific and Medical (ISM) frequency band (2.4 GHz). To minimize potential interference with other technologies using this crowded radio band (WiFi, microwave ovens, etc.), Bluetooth employs a frequency hopping technique (Bensky, 2007). For technical details about the Bluetooth protocol and the discovery and communication process, the reader is referred to Peterson et al. (2006). In brief, the process can be summarized as follows. To set up a Bluetooth network ('piconet') in which all devices follow the same hopping frequency to communicate with each other, a Bluetooth device first needs to be discovered by a master device during an inquiry phase. This master device enters the 'inquiry' substate and starts transmitting inquiry packets, which triggers unknown devices within the detection range of the master device that are in the 'inquiry scan' substate ('discoverable') to respond by transmitting an identification message containing their MAC address and class of device (COD) code. In a next phase, called the page phase, the master and slave devices synchronize their internal clocks to align their frequency hopping schemes and start communicating.

All Bluetooth scanners used in this study functioned as master devices, continuously broadcasting inquiry messages and listening for the responses of other discoverable devices that passed within their radio range. This way, mobile devices were detected without the need for an actual connection with the scanner. All further processing was done in the scanner. In comparison with other studies that did need a connection (Hay and Harle, 2009), the methodology used in this experiment is better suited for mass events where the participation of the tracked individuals should be minimal. Every time a device was detected, its MAC address, COD code, and the timestamp of the detection were registered. Additionally, the received signal strength intensity (RSSI) of the inquiry response was registered. This intensity value is inferred from the received power level with which the response packet was detected by the scanner and is theoretically negatively correlated with the distance between the scanner and the detected device (Hossain and Soh, 2007). Although it is technically possible to also register the 'friendly name' of the device (which can be changed by the users), this would significantly slow down the scanning process. Additionally, part of the tracked population might be identified by the inclusion of names or phone numbers. As this would raise serious privacy issues, this information was not registered.

In theory, the RSSI values registered by different scanners can be used to calculate the position of a mobile device through multilateration (Bensky, 2007; Kelly, 2010). Although some studies have reported acceptable accuracies in indoor environments, they have also outlined problems with multipath fading due to obstructions in these environments (Feldmann et al., 2003; Zhou and Pollard, 2006). Because of the complex environmental setting and the resulting unpredictability of the propagation of Bluetooth signals, positioning in this case study was done through the 'proximity' principle, where the position of a detected mobile device is approximated to the position of the sensor by which it is detected (Bensky, 2007). This way, the path of a mobile device is reconstructed by means of a sequence of time intervals spent within the detection ranges of different Bluetooth scanners. The precision of the resulting trajectories ultimately depends on the detection range of the Bluetooth scanners, and on their number and distribution over the study area. In theory, Bluetooth devices are classified into three power classes. Class 1, 2 and 3 devices support theoretical ranges of 100 meters, 10 meters and 1 meter respectively. The observed detection range, however, depends on how many obstructions the signal encounters between the scanner and the mobile device (buildings, furniture, clothing, people, etc.).

2.2.2 Equipment

Figure 2.1 on the following page shows the hardware components used in this Bluetooth tracking experiment. A Bluetooth 'scanner', as referred to in the remainder of this chapter, is a combination of a computational unit running the scanning software and processing and storing the results (1), a power source (2), and a USB cable (3) to connect the computational unit with a Bluetooth sensor (4–5). The heart of the computational unit was an ALIX motherboard (alix2d2, alix3d2), equipped with a 1 Gb CompactFlash card for storing log files. The operating system was an adapted version of *Voyage Linux* (based on *Debian* 5, 2.6.32.15 kernel), and the scanning software, *Gyrid*[1] (version 0.3.3), was developed at our research group. It is a python implementation built around the *BlueZ* Bluetooth stack (version 4.63).

In order to control the detection range, different Bluetooth sensors were used. In this study, we used a combination of class 2 (D-Link DBT-122) and class 1 (Sena Parani UD-100) devices. These are respectively shown as numbers 4 and 5 in figure 2.1 on the next page. At most locations, class 2 devices without an external antenna were used. Where a larger detection range was necessary, a class 1 device with a replaceable antenna was used. Two types of omnidirectional antenna were available with gains of 3 dB_i (6) and 5 dB_i (7). The higher the gain of the antenna, the higher the detection range was. Experiments showed that the detection ranges deviated considerably from the theoretical ranges in case of line-of-sight detections (up to 100 meters for the class 2 device, and 300 m for the class 1 device with a 5 dB_i external antenna).

[1] http://github.com/Rulus/Gyrid

Figure 2.1: Bluetooth hardware used in the tracking experiment: computational unit (1), power source (2), USB cable (3), class 2 Bluetooth dongle with internal antenna (4), class 1 Bluetooth dongle (5), and 3 dB_i (6) and 5 dB_i (7) omnidirectional antennas.

2.3 Background and experimental design

2.3.1 Description of the event and study area

The 'Ghent Festivities' is a ten day long cultural and theatre festival taking place in the historic city centre of Ghent in Belgium. Besides stage events at the larger public squares (starting in the evening), there are also random small street acts during the day resulting in an almost continuous flow of events. The festival has grown from a small event in the sixties attracting only local residents, to a genuine mass event attracting visitors from larger distances. Because of the size and the openness of the event — most activities in the festival are free, and there are no explicit entrance or exit points — collecting quantitative data such as visitor counts is challenging. The resulting lack of quantitative data acts as a bottleneck for research into the spatiotemporal dynamics of visitor movements. Exemplary to this is the issue of calculating the total number of visitors that attend the festival, which has traditionally been estimated by using proxy variables such as the daily amount of waste collected in the centre and the number of tram or bus tickets sold (Jansen-Verbeke et al., 2003). As such, estimations vary but the general consensus is that 1.5 million (non-unique) visitors attended the festival in 2010. Other than this rough figure and the use of video technology by the police department to give a qualitative indication of crowdedness, little is known about the general movement patterns of these visitors within and around the festival site, how long they stay at the festival, the number of days they visit the festival, which transport mode they use in order to reach the festival, etc. This chapter aims to highlight the potential of Bluetooth tracking for understanding such crowd dynamics, and thereby significantly complement the (little) quantitative data that is currently available.

In order to manage such a large event in a relatively small area, a body of regulations has been prescribed by local authorities and other stakeholders (police department, fire department, festival organisers and residents). Specific rules apply in the officially documented 4.5 km^2 festivities zone, surrounding the historic part of the city centre. This zone comprises 11 public squares which act as major attractors because of on-stage performances, bars, food stands, fairs, etc. Because of the large size of the event, mobility issues regarding the movement of visitors to and from the event are also important. Consequently, the study area was defined larger than the festivities zone *sensu stricto*, and we also focused on the two main train stations in Ghent, and a park&ride facility in the southwest. In this park&ride facility visitors could park their car and subsequently take a tram to the city centre. A general overview of the study area is depicted in figure 2.2 on the following page. A summary of the different locations that were covered with a Bluetooth scanner is given in the next section.

2.3.2 Selection of scanner sites

Given the limited range of the Bluetooth scanners and the size of the event, a full coverage of the entire study area was impossible from a practical point of view. Instead, a careful selection of strategic coverage sites was made after consultation with local policy makers and urban experts with the purpose of collecting as many significant individual movements as possible. The spatial distribution of the selected locations is depicted in figure 2.2 on the next page. In total, 22 locations were covered. In order to capture the main bulk of movements between the 11 public squares, a scanner was placed at each square (scanners 1–11). A selection of points of access around the festivities zone was also covered (scanners 12–19). One of these was located in the busiest shopping street of Ghent (scanner 12). The two main train stations of Ghent were also covered with a Bluetooth scanner at locations where the majority of train passengers needed to pass by. Finally, the tram station next to the park&ride facility in the southwest was covered by two scanners in order to be able to track all passengers. In the rest of the analyses, these two scanners are regarded as one scanner by merging all detections. All scanners were operating for the entire duration of the event, except for scanners 13–19 (points of access) which were only operational during the two first days of the festival. As mentioned above, the type of sensor was chosen in function of the desired detection range at each site. Inside the festivities zone, the scanners in the smaller squares (scanners 5, 6, 8, 9 and 11) were equipped with a class 2 Bluetooth dongle and the scanners in the larger squares (scanners 1, 2, 3, 4, 7 and 10) with a class 1 Bluetooth dongle. The type of antenna for the latter dongle was chosen in function of the size of the square that needed to be covered (the larger the area, the higher the gain). As

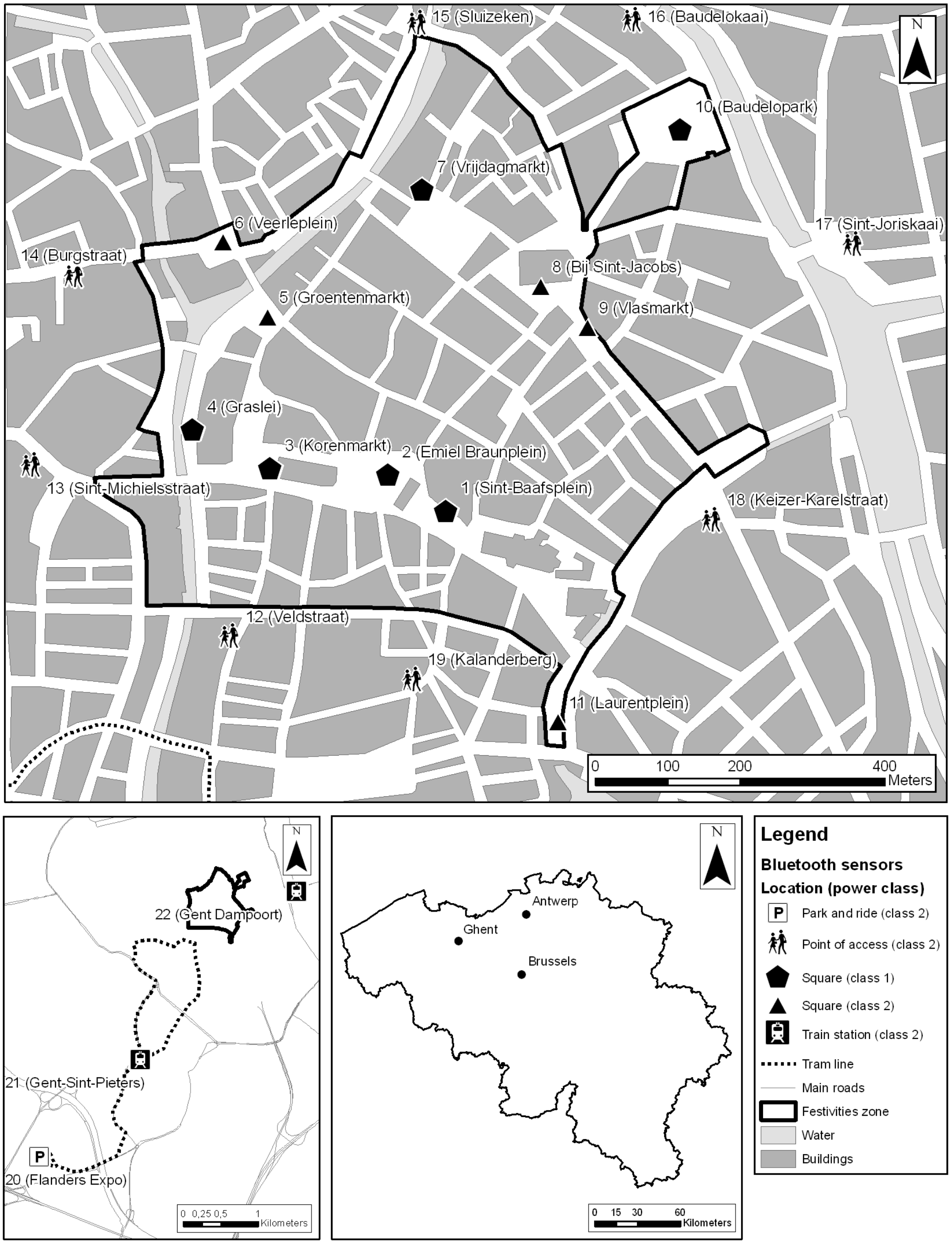


Figure 2.2: Overview of the study area and location of Bluetooth scanners.

such, scanners 1, 2, 3, 4 and 7 used a 5 dB_i external antenna, and scanner 10 a 3 dB_i external antenna. All scanners outside of the festivities zone used a class 2 Bluetooth dongle.

2.4 Preprocessing

The raw data consisted of log files on the different scanners having the following format: timestamp of detection, MAC address of the detected device, COD code of the detected device, RSSI of detection. After merging the log files of the different scanners, the dataset consisted of 263,680,889 loglines. In order to obtain a compressed dataset, the scanners were programmed to create a second set of log files during the scanning process in the following compressed format: timestamp, MAC address of the detected device, COD code of the detected device, ‘in’/‘out’/‘pass’. A buffer time of 10 seconds was used to create detection time *intervals* from the detection time *points*. ‘In’ was written when a device entered the detection range of the sensor, and ‘out’ was written when the device left the range. ‘Pass’ was used for solitary detections with no prior or later detections within 10 seconds. This way, the dataset was compressed by 91% to 23,889,850 loglines. Figure 2.3 shows an extract of both types of logged data. Over the entire duration of the event, 102,467 unique devices were detected over all covered locations. The majority of these were detected at least once inside of the festivities zone (88,763 devices or 87%). The remainder of the analyses were done using a ‘geographical information system for moving objects’ (*GISMO*), implemented in *Java* and developed at our research group.

```
20100720-175338-CEST,20:21:A5:45:40:40,5898756,-81
20100720-175340-CEST,20:21:A5:45:40:40,5898756,-80
20100720-175341-CEST,20:21:A5:45:40:40,5898756,-72
20100720-175353-CEST,20:21:A5:45:40:40,5898756,-78
20100720-175355-CEST,20:21:A5:45:40:40,5898756,-82
```

$\downarrow$

```
20100720-175338-CEST,20:21:A5:45:40:40,5898756,in
20100720-175341-CEST,20:21:A5:45:40:40,5898756,out
20100720-175353-CEST,20:21:A5:45:40:40,5898756,in
20100720-175355-CEST,20:21:A5:45:40:40,5898756,out
```

Figure 2.3: Extract of logged data showing the raw time point detection data (top) and the compressed time interval data (bottom). This example shows one Bluetooth device (MAC address `20:21:A5:45:40:40`) being detected 5 times on 20/07/2010 between 17:53:38 and 17:53:55 (CEST: Central European Summer Time). The buffer time of 10 seconds causes the raw data to be split into two separate detection time intervals (in $\rightarrow$ out). The COD code of the device (5898756) shows that this was a regular cell phone.

Because this study uses mobile devices as a proxy for detecting the movements of their mobile users, we have analysed the different types of devices that were detected by the Bluetooth scanners. As mentioned above, detectable devices respond to the inquiry call of the scanners by transmitting their class of device code. This code directly corresponds to certain types of devices. The scheme that is thereby followed was created by the Bluetooth SIG (Special Interest Group) and distinguishes between six major classes (‘Audio/Video’, ‘Computer’, ‘Imaging’, ‘Network Access Point’, ‘Peripheral’ and finally ‘Phone’). Additionally, some devices do not publish their COD (‘Unknown’). The minor classes give more information about the specific kind of device within each major class. They distinguish, for example, between cell phones and smartphones, laptops and desktops, etc. Taking both major and minor classes into account, the dataset was divided into three general kinds of devices: phones (91%), handsfree car kits (7%) and other devices (2%). The car kits represent cars, while the phones represent persons. The other devices devices could not be directly linked to a (moving) person. Because this study focuses on the movements of persons, only phones were withheld in the remainder of the analyses. The resulting dataset consisted of 80,828 phones or visitors that were detected at least once in the festivities zone during the event.

Because the ‘Ghent Festivities’ span multiple days, some of the visitors visit the event on more than one day. Consequently, there is a need for a distinction between *visitors* (the person itself modelled as a moving object with a fixed identifier) and their *visits* (modelled as trajectories generated by the moving object). Because of the continuous flow of events, a fixed time point at which to split trajectories was not available. Accordingly, an algorithm using a ‘maximum time gap’ parameter was developed. Trajectories containing gaps between detections with a duration of more than this maximum time gap were split into two subtrajectories. After careful exploration of the dataset, this parameter was set at 5 hours. As a consequence, the 80,828 detected phones/visitors that were detected at least once inside of the festivities zone were responsible for 152,487 trajectories/visits in total.

To make predictions about the entire visitor population, observed numbers of detected phones/trajectories need to be extrapolated to estimated numbers of visitors/visits. In order to do this, the ‘detection ratio’ (the percentage of visitors that gets detected by means of a mobile device with a visible Bluetooth interface with respect to the entire population) needs to be known. To this end, we compared visual counts of passing people in smaller passageways with the number of Bluetooth devices detected during the same time period by a mobile scanning setup (laptop + class 2 Bluetooth dongle). In order to preserve privacy, we did not record any images in this process. Ten such experiments were conducted at eight different locations, each of them over a time period of 15 minutes. Averaging the results yielded a general estimation of the detection ratio of 11.0$\pm$1.8%. This ratio can be used to roughly estimate total numbers of visitors/visits from numbers of detected devices/trajectories. Care should be taken, however, when generalising insights gathered from tracked visitors to the entire visitor population. This is due to the possibility of sample bias resulting from different detection ratios in different population segments, which is further discussed in section 2.6.3 on page 39.

2.5 Results

In the remainder of this section, we will highlight some of the analytical possibilities of Bluetooth tracking data in the context of mass events by showing a selection of concrete results from the ‘Ghent Festivities’.

2.5.1 Total number of visitors and visits of the event

Given the detection ratio of 11.0±1.8%, we can estimate the total number of unique visitors and visits (or 'non-unique' visitors). This leads to an estimation of around 735,000 unique visitors (minimum: 630,000, maximum: 880,000) and around 1.4 million visits (minimum: 1.2 million, maximum: 1.7 million). The estimation of 1.5 million visitors made by the city department is apparently in line with our estimations.

2.5.2 Total number of unique visitors per location during one day

The aggregated number of phones that is detected over a certain time interval on a Bluetooth scanner gives an indication of the total number of visitors that passed within the detection range of a sensor during that time interval. Because of the strategic locations of the Bluetooth scanners, the tracking data can be used to estimate the number of visitors that visit each location in and around the festivities zone. This is illustrated in figure 2.4, which shows the number of phones detected at each public square during the third day of the event. It is immediately clear that some squares attract considerably more visitors than others. Square 3, for example, attracts more than three times as many visitors than square 11 (7,639 vs. 2,347 detected phones) over the course of a day. In general, the public squares in the south-west of the festivities zone attract most visitors, reflecting the fact that the largest and most popular squares are situated there as well as the main points of access. By multiplying these counts with the previously defined detection ratio, it is possible to make a rough estimation of the total number of visitors that visited each location. The busiest square for example attracted around 70,000 visitors on that specific day.

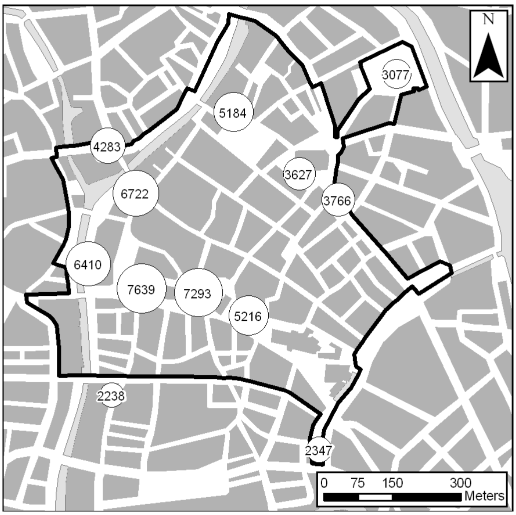


Figure 2.4: Aggregated number of detected phones on the third event day of the 'Ghent Festivities' (19/07/2010 11 am until 20/07/2010 7 am) at the 11 public squares and the main access point (12).

2.5.3 Varying number of visitors in the entire festivities zone over time (day, hour)

By aggregating the number of detected phones over regular time periods, it is possible to calculate the crowdedness at a location over time. Additionally, a number of locations can be generalised into one zone to estimate zonal crowdedness. For this case study, we calculated the crowdedness in the festivities zone at time resolutions of one day and one hour. This is illustrated in figure 2.5 on the next page. Aggregating over one hour time windows results in a very smooth curve with sharp troughs in the morning (usually around 7 am). The peaks are also usually sharp and situated around 11 pm except for days 2, 5 and 9 where a broader peak in the late afternoon is observed. These correspond to two Sundays and the national day of Belgium (21 July), and these days are known to attract more daytime visitors (such as working couples with children). As a result, the sharp peaks around midnight do not appear because of the relatively larger crowdedness earlier in the afternoon. The three busiest days are immediately visible: the fourth day is the most crowded with almost 10,000 detected phones or 90,000 unique visitors in the festivities zone between 11 and 12 pm. The little peak during the build-up of the first day is due to the opening parade on the first day attracting specific visitors leaving immediately afterwards. To aggregate over daily periods, one should carefully consider how to define a day. Looking at the hourly crowdedness, it does not make sense to define days starting and ending at midnight because that is generally the most crowded period of the day. Doing so would cause the Bluetooth observations to be segmented by unnatural breaks. Consequently, we have considered the starting point of an 'event day' to coincide with the on average least crowded moment of a day, i.e. 7 am. The daily aggregates again show the three busiest days with day 4 peaking at almost 20,500 detected phones or 190,000 visitors.

2.5.4 Changing distribution of the crowd in the festivities zone over time

Although the varying crowdedness in the festivities zone already offers valuable insights for all stakeholders of the festival, it does not take the spatiotemporal dynamics of the crowd within the zone into account. As a first and general approach to shed light onto these dynamics, we have examined the changing distribution of the crowd over the different public squares in the festivities zone over time. This was done by slicing the data into time periods of one hour, and counting the number of phones detected at each public square during each time period. Subsequently, these numbers were summated over the ten event days. The resulting chart in figure 2.6 on page 32 shows the changing distribution of the crowd in the festivities zone during a generalized event day. A clear trend is visible where the crowd is evenly distributed over the entire centre for most of the day, but condenses to a smaller set of squares during the night. Especially locations 8, 9 and 10 seem to attract a specific night audience (together they attract around 50% of the entire audience between 5

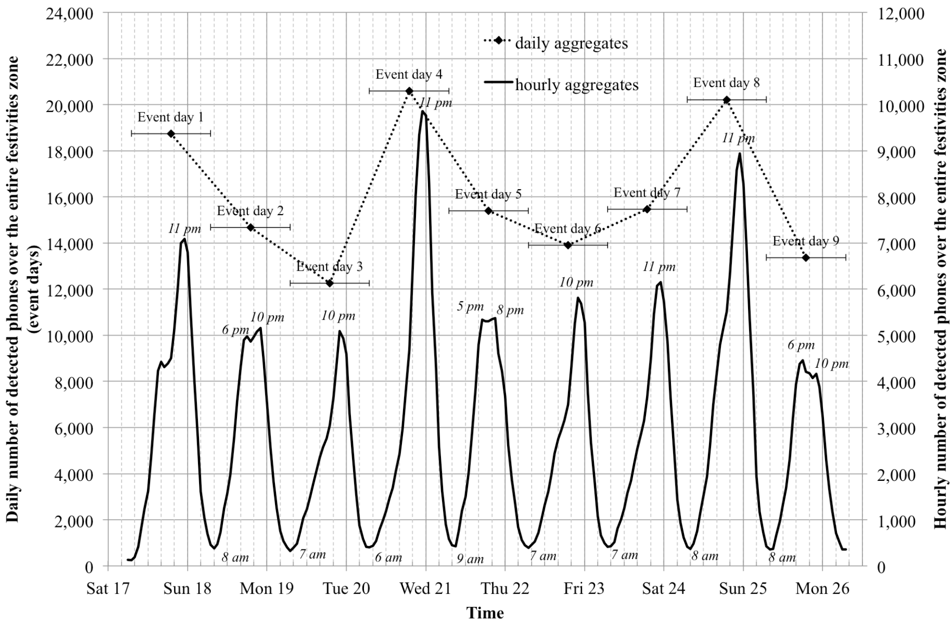


Figure 2.5: Daily (event days starting and ending at 7 am) and hourly number of detected phones over the entire festivities zone as an indicator of crowdedness. Solid vertical gridlines point to midnights, dashed vertical gridlines are plotted every 4 hours.

and 6 am). This well-known phenomenon is caused by visitors gathering in this area after midnight and staying until the morning. Although it is not shown in this figure, this trend was visible during every day.

2.5.5 Returning visitors

Because of the fixed MAC address assigned to each device, it is possible to investigate how many detected visitors visit the event for more than one day. This was done by counting the number of trajectories (visits) per phone (visitor). If a trajectory comprised more than one day, only the first day was considered to prevent potential errors from visitors staying longer than 7 am and being registered as two-day visitors. The result is shown in figure 2.7 on the next page. There is a dominance of one-day visitors with respect to returning visitors (65% vs. 35%). Furthermore, the share of returning visitors decreases with an increasing number of visit days (from almost 14,000 phones for 2 days to 131 phones for 10 days).

The share of returning visitors at each public square *separately* was calculated to gain insight into the tendency of returning visitors to visit the same squares in the festivities zone. The used procedure was the same as described in the previous paragraph. The result is depicted in figure 2.8 on page 33. Squares 9 and 10 are characterised by a higher than average degree of returning visitors (21.6±0.2% for these two squares vs. 13.5±3.4% for the other

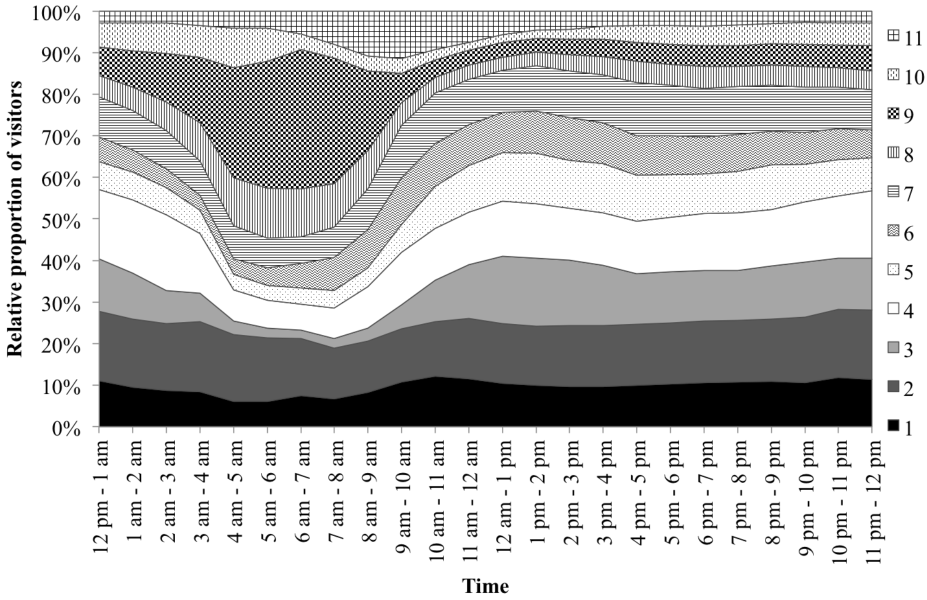


Figure 2.6: Distribution of the detected crowd over the different public squares in the festivities zone over time (hourly aggregates, summated over the ten event days).

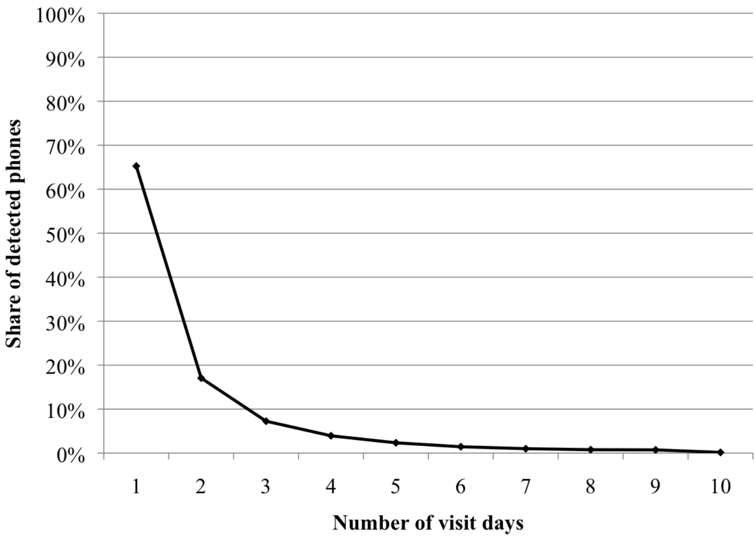


Figure 2.7: Share of detected phones in function of the number of visit days.

squares), whereas locations 5 and 11 exhibit a lower degree. The anomalous observations at locations 9 and 10 can again be attributed to their functionality as night hubs of the event.

Apparently, visitors of this area seem to return for more than one day more often than the average visitor at other locations. Square 5 is a smaller square in the centre of the festivities zone, and a significant portion of the more general public (consisting largely of one-day visitors) needs to pass this location to walk around the centre. The result at square 11 is influenced by its proximity to the border of the festivities zone and hence to movements that are not related to the event.

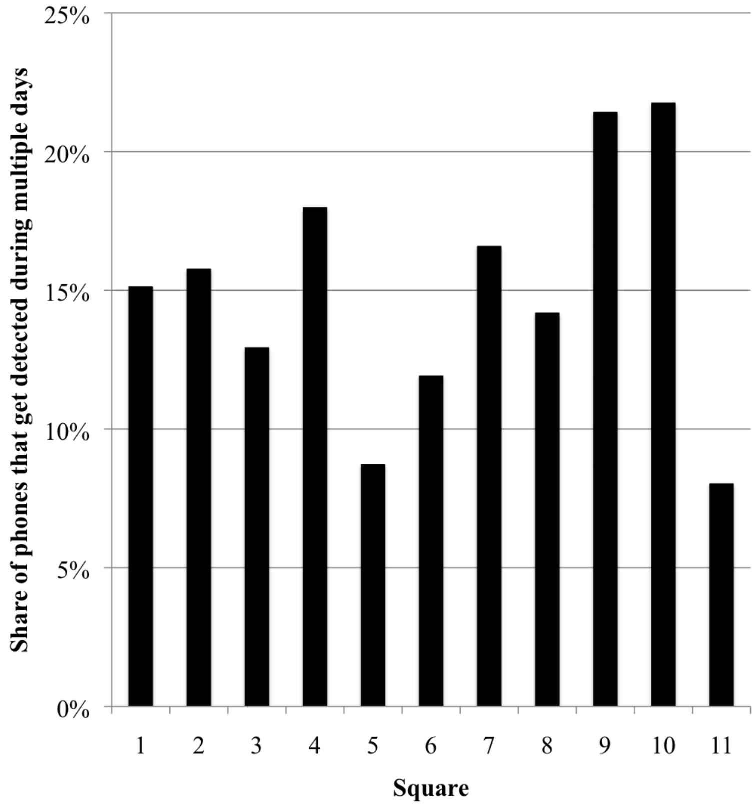


Figure 2.8: Share of phones detected on more than one day at the different public squares inside of the festivities zone.

2.5.6 Transportation mode

Bluetooth tracking data can additionally offer information on transportation modes by a careful selection of sensor locations: train users can be distinguished in a train station, tram users in a tram stop, car users in a parking lot, etc. In this case study, train users were detected by scanners in the two train stations of Ghent. Visitors making use of the park&ride facility outside the centre were detected at the main tram stop upon entering the tram taking them to the centre. Users of other transport modes were harder to track because of a lack (pedestrians, cyclists) or an excess (car and bus users) of fixed departure points. Despite the fact that not all transport modes were detected, these two groups do represent visitors coming to Ghent from larger distances. As an illustration, we calculated their relative shares within the total visitor population for the different days of the event. The result is shown in figure 2.9. The relative share of tram users clearly varies more over time (3–7%) than the share of train users (5–6%). Remarkably, the share of tram users follows the exact same trend as the daily visitor counts in figure 2.5 on page 31. Apparently, there is a systematically larger share of visitors making use of the park&ride facility and going to the city centre by tram on the busier days. This is interesting as it could indicate a change in composition of transport modes of the public over the different days.

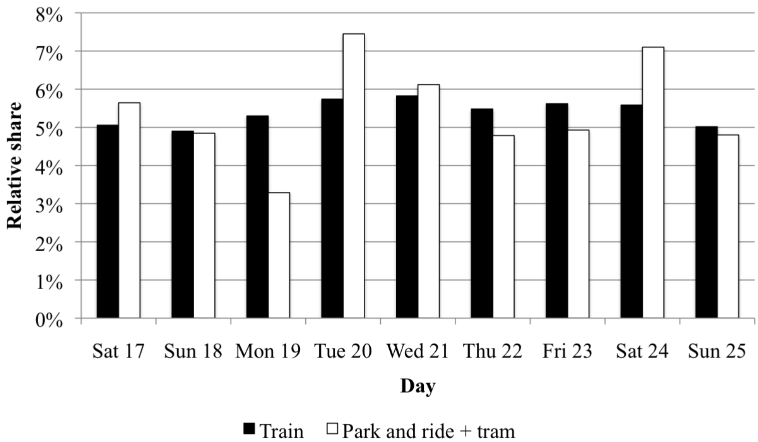


Figure 2.9: Relative share of train and park&ride users on the different days of the event.

2.5.7 Visit duration

Technically, there are two methods of estimating the duration of a visit to the 'Ghent Festivities'. The first method involves measuring the total duration of detection by generalising the detections over the different public squares into one zone. Because of the incomplete coverage of the study area, however, gaps occur in between the detections. Algorithms can merge all co-located detection intervals within a certain time threshold of each other, but the correct choice of this threshold is problematic without accurate and systematic comparisons with ground truth. Consequently, we chose to estimate the duration of a visit in a second way. Because some access points around the festivities zone were covered during the first two days, the duration between two detections at an access point separated by one or more detections in the festivities zone can also be used as a proxy for the duration of a visit. The resulting distribution in figure 2.10 on the next page exhibits a wide spread around its median value of 3.5 hours. The majority of this subpopulation of visitors seems to visit the event for only a few hours, while the large positive skew shows that other visitors stay much longer (1,036 trajectories or nearly 11% of the sample stay for at least 7 hours). Because there seems to be a large variability in visit duration among (certain types of) visitors, the median value should be regarded as a very rough generalisation of this distribution.

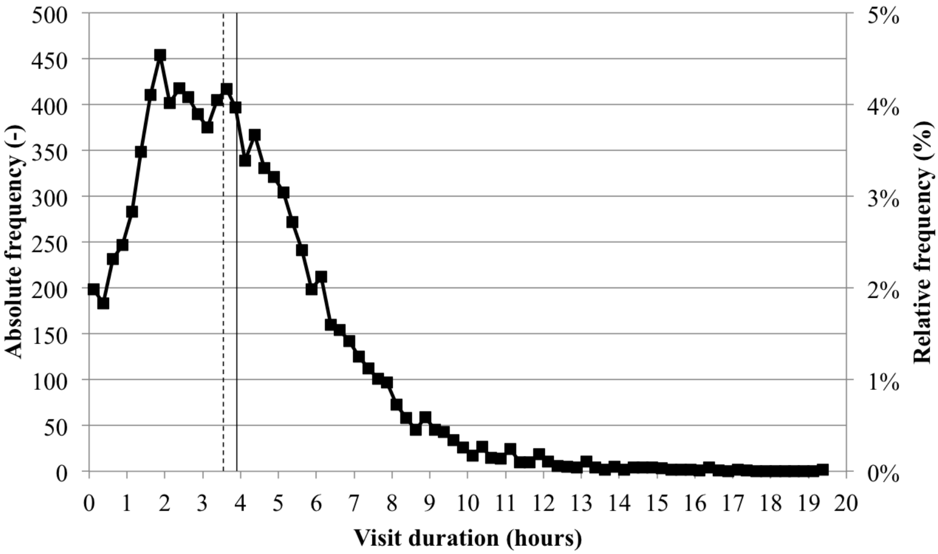
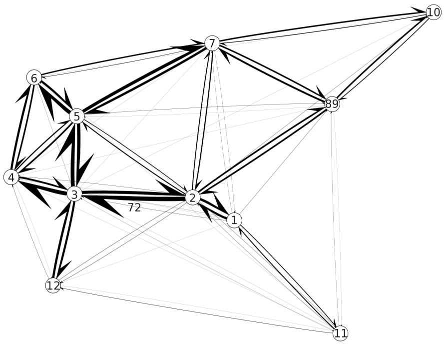


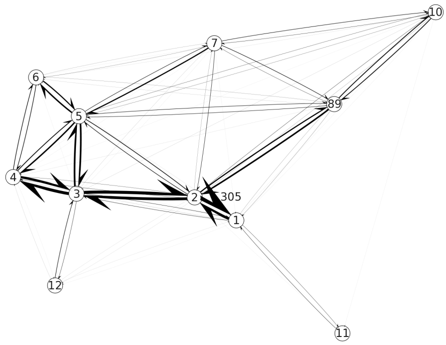
Figure 2.10: Histogram of the duration of a visit to the 'Ghent Festivities' (class-width: 15 min, sample size: 9,648). The average value is depicted by the solid line (3 hours, 53 minutes and 58 seconds), the median value by the dashed line (3 hours, 32 minutes and 42 seconds).

2.5.8 Flow analysis

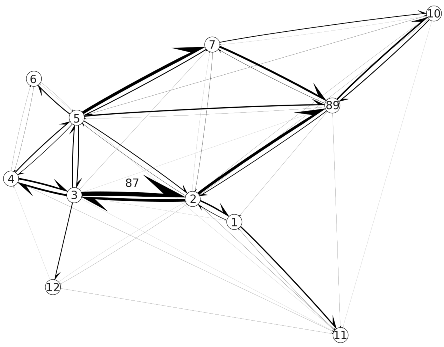
Although our methodology can only analyse flows of visitors carrying discoverable devices, the discovered patterns and trends can aid stakeholders to make well-informed decisions regarding crowd-management and security in general. By making a time series of these flow diagrams, it is possible to investigate the time-dependency of visitor flows and link them to factors that potentially influence the movements such as the time and order of performing artists at the different locations. In this chapter, we limit the flow analysis to the third event day (19/07–20/07) and study the flows without an in-depth look at their influencing factors. A flow from location A to location B is defined as the number of mobile devices that is subsequently detected at A and B over a maximum time period of 30 minutes, without being detected at any other locations in between. Figure 2.11 on the following page shows four characteristic snapshots of the visitor flows during this event day. Regular time periods of 30 minutes were used to generate the different snapshots, and flows were attributed to their respective time period according to their departure time. The first snapshot (a) shows that the majority of the flows in the afternoon take place in the southwest of the festivities zone. A clear west-east and north-south axis is visible. Later at night (b), movements seem to condense along the west-east axis. The large flow from location 2 to 1 — which was also the largest flow of the entire event day — is related to the performance of a popular artist at location 1 at 11 pm which temporarily caused square 1 to be closed to prevent overcrowding. Around 3 am (c), flows pointing to the night-hub in the northeast (squares 8, 9 and 10) dominate. Later in the morning (d), there are still considerable flows to this area but flows moving back to the southwest dominate more. Since there are no programmed events at this time in the southwest, these flows largely consist of visitors leaving the event.



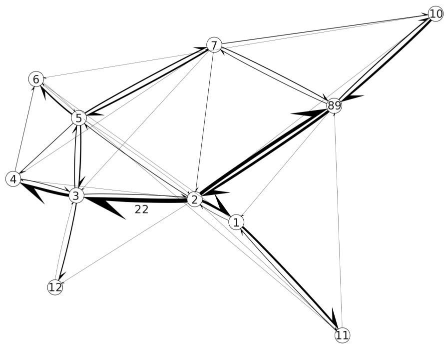
(a) 19/07/2010 4 pm–4:30 pm



(b) 19/07/2010 10 pm–10:30 pm



(c) 20/07/2010 3 am–3:30 am



(d) 20/07/2010 4:30 am–5 am

Figure 2.11: Four snapshots of visitor flows in the festivities zone during periods of 30 min. The direction of the arrow indicates the direction of the flow, the width of the arrow indicates its size. The widths of the arrows are normalised to the size of the largest flow during each time period separately (indicated by the number next to the widest arrow). The numbers in the circles refer to figure 2.2 on page 26. Locations 8 and 9 were generalised into one zone '89' for easier representation. Movements taking less than 30 seconds or longer than 30 minutes were discarded, and the remaining movements were allocated to the time period containing their departure time.

2.6 Discussion and conclusion

The discussion is organised as follows. First, the added value of Bluetooth tracking for mass events is discussed by reviewing the results generated for the 'Ghent Festivities'. Next, we describe why Bluetooth has a high potential as a tracking technology in this specific niche of use-cases. We conclude by highlighting some remaining issues and weaknesses concerning the methodology, and make suggestions for further research.

2.6.1 The added value of Bluetooth tracking for mass events

We were able to generate a huge dataset of movement trajectories with a limited number of scanners. Around 100,000 devices were detected over ten days. This is equal to the amount that was detected by 450 nodes over four months in the 'Cityware' project, based in the United Kingdom (Kostakos and O'Neill, 2008), and about 5 times higher than the amount detected by 16 scanners over two days in the 'Donauinselfest' (Leitinger et al., 2010). Comparing the total number of detections in the raw data with the number of detected devices, we can see that each device was on average detected 240 times by one or more of the scanners. This is a first and strong indication that the employed technology and system are able to manage the huge flow of information generated by tracking visitors of very large mass events such as the 'Ghent Festivities'.

In order to determine the true quality of the gathered tracking data, the extracted information needs to be confronted with reality. One such figure is the number of devices/ trajectories in the database as representatives of visitors/visits, which can be estimated by means of a detection ratio. Our detection ratio estimate of 11% seems higher than the ones of 7% found in two other studies (O'Neill et al., 2006; Weinzerl and Hagemann, 2007). The increasing penetration of the Bluetooth technology in recent years might be responsible, but ultimately there is a need for more automated and accurate ratio estimations if Bluetooth tracking results are to be extrapolated to an entire population of moving individuals. Nevertheless, the estimated number of visits calculated from our dataset seems to closely correspond to the estimation made by the city department (1.4 vs. 1.5 million respectively). This indicates that the detection ratio is reasonably accurate and that there is no significant problem with devices somehow escaping the detection ranges of our scanners (O'Neill et al., 2006).

Both the daily and hourly variations of crowdedness in the festivities zone showed a smooth profile with recurring peaks and troughs. The crowd density profiles for the separate public squares (not shown) followed similar patterns, highlighting the suitability of Bluetooth tracking for studying spatiotemporal crowd density variations at specific locations or zones. Apart from distinguishing between crowded and less crowded days, we were also able to identify three days characterised by a relatively larger share of afternoon visitors.

Large events such as the 'Ghent Festivities' are known to attract a wide spectrum of visitors with different tendencies and expectations. This is noticeable in the rest of the analyses. While the public squares in the southwest of the festivities zone attract the main share of visitors (Figure 2.4 on page 29), the squares in the northeast clearly function as a night-hub of activities (Figure 2.6 on page 32). Additionally, visitors of this night-hub seem to return more regularly than visitors that avoid this area (Figure 2.8 on page 33). Despite the clear dominance of one-day visitors (Figure 2.7 on page 32), the large size of the event makes that the relatively small share of visitors that attend the event for several days nevertheless constitutes a group of considerable size. As an example, 2,173 phones were detected during at least seven days, which represents a group of around 20,000 visitors.

It was possible to make an estimation of the average visit duration, but the resulting distribution (Figure 2.10 on page 35) again pointed to the existence of two (or more) types of visitors: a majority that only visits the event for a few hours and a relatively smaller group that stays considerably longer. As a result, the median value of 3.5 hours does not reflect all visitors and probably lies closer to the average visit duration of the majority of the public than that of the longer staying visitors.

Because of the coverage of two train stations and a large park&ride facility outside the centre, the transport mode of roughly 10% of the visitors could be determined (either train or tram after parking the car). A preliminary analysis showed that the share of visitors using the park&ride facility was disproportionally higher on days where more visitors in general attended the event. This could reflect a temporal change in composition of transport modes, but more research and a higher coverage of transport modes are necessary to confirm this hypothesis.

A concise flow analysis exposed a typical spatiotemporal pattern over the course of one event day that was in agreement with on-field experiences of the city and police department. Together with qualitative estimations of crowd density by means of video cameras distributed over the festivities zone, they are especially valuable for assisting crowd-management specialists.

2.6.2 The 'Bluetooth niche'

If a tracking technology is to be suitable for studying spatiotemporal dynamics of crowds at mass events, it needs to comply with certain requirements. In this section, we explain why Bluetooth tracking can potentially fill this niche by comparing with two other technological candidates for this use-case: GPS and mobile positioning of cell phones (Ahas et al., 2008; González et al., 2008).

First and foremost, Bluetooth tracking offers the ability to track a very large number of individuals in an easy and relatively inexpensive way. A large population sample is often necessary to understand the complex dynamics involving crowd movements at mass events. Other tracking technologies, such as GPS, only reach a small subset of individuals because of the low penetration rate in the general audience (Ratti et al., 2006). Additionally, they involve the tracked individual in a direct way (e.g. by distributing and later recollecting logging units). This renders the tracking process labour-intensive, and makes it prone to possible bias since individuals might behave differently because they are aware of being tracked or because certain population segments might be more inclined to cooperate with such experiments and hence be over-represented in the resulting dataset. Mobile positioning datasets are usually even larger than Bluetooth tracking datasets, but cooperation with mobile operators has proven to be difficult (Ahas et al., 2008). The ultimate cost of the technology will depend on the number of scanners (around € 200 per device) used, but will generally be fairly low if expressed per tracked individual.

Secondly, Bluetooth tracking has a clear advantage over other technologies because of its ubiquitous applicability in indoor as well as outdoor environments. Because of the freedom with which Bluetooth scanners can be placed in indoor environments, it is possible to follow individuals at room level inside buildings. Signal deterioration makes this practically impossible with GPS, and very challenging with mobile positioning.

Finally, the limited detection range of Bluetooth sensors makes the resulting trajectories more accurate than those in mobile positioning datasets. As mentioned above, the actual range of our Bluetooth scanners varied from around 10 to 100 meters, depending on the used sensor. Location estimations in mobile positioning data typically have an accuracy of a few hundred metres in urban settings (Ahas et al., 2007), which is insufficient for distinguishing between locations that are less than 100 meters apart (such as locations 8 and 9 in the festivities zone).

2.6.3 Remaining issues and suggestions for future research

Although we have demonstrated that Bluetooth tracking is able to deliver valuable information to stakeholders of mass events, there are some remaining issues that need to be addressed. One of the prime issues is the possibility of biased results by over- sampling certain segments of the total population of individuals (Rice and Katz, 2003). Adolescents with a higher education might indeed carry more Bluetooth-enabled devices than elderly people, while young children will probably never be detected. The potential difference in Bluetooth usage among different audiences might significantly influence generated insights. Accordingly, more research is needed into the use of discoverable Bluetooth-enabled devices by different population segments in order for Bluetooth tracking to evolve into a technology delivering accurate and reliable information to policy makers. The estimated detection ratio of 11.0±1.8% is sufficient for making rough extrapolations to the entire population of visitors, but a more systematic way of calculating the percentage of the population being tracked will be necessary for more reliable extrapolations in the future. Additionally, the possible influence of time and space on the detection ratio needs to be investigated.

Focussing on this case study, several analyses have exposed the existence of multiple profiles of visitors. More research is needed into which profiles exist, and how they can be distinguished from each other. This could either be done with the current dataset containing only locational information (e.g. the time that a visitor spends on a certain public square might be used to define different visitor profiles), or with a new dataset enriched with socio-economical information about the tracked individuals. In the latter case, we might statistically prove the likely assumption that there are two main visitor profiles (adult visitors in the afternoon and adolescents during the night) and study the influence of socio-economical attributes on movement behaviour.

Transportation mode detection was adequate for train and tram users, but the scope should be extended to include other transportation modes as well (cars, pedestrians, cyclists). More research is needed into the influence of the transportation mode on visitor behaviour, such as the duration of a visit or the movement patterns between the public squares.

The tentative flow analysis already suggested certain patterns, but further research should be devoted to finding representative patterns in the order in which the different squares are visited. Sequence alignment methods seem the most likely candidate for extracting these patterns (Shoval and Isaacson, 2007a; Wilson, 1998) but they might need further modifications to handle the spatiotemporal complexity of Bluetooth tracking data gathered at large mass events such as the ‘Ghent Festivities’.

Because of the difficulty of directly correlating the RSSI of a detection with the distance between the sensor and the detected device, we have used the proximity principle to generate trajectories from the detection data. Fine multilateration of the signal strengths registered on different sensors to calculate an accurate location seems unrealistic, but a rough multilateration might be possible under certain circumstances. In this way, a continuous crowd density over a public square might be calculated while this is not possible using the proximity principle.

References

Ahas, R., Aasa, A., Mark, U., Pae, T., and Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898–910.

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Ahlqvist, O., Ban, H., Cressie, N., and Shaw, N. Z. (2010). Statistical counterpoint: Knowledge discovery of choreographic information using spatio-temporal analysis and visualization. *Applied Geography*, 30(4):548–560.

Andrienko, N. and Andrienko, G. (2007). Designing Visual Analytics Methods for Massive Collections of Movement Data. *Cartographica*, 42(2):117–138.

Axhausen, K., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.

Batty, M., Desyllas, J., and Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, 17(7):673–697.

Bensky, A. (2007). *Wireless positioning technologies and applications.* Artech House, Boston, London.

Bogorny, V., Kuijpers, B., and Alvares, L. O. (2009). ST-DMQL: A Semantic Trajectory Data Mining Query Language. *International Journal of Geographical Information Science*, 23(10):1245–1276.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Bullock, D. M., Haseman, R., Wasson, J., and Spitler, R. (2010). Anonymous Bluetooth Probes for Measuring Airport Security Screening Passage Time: The Indianapolis Pilot Deployment. *Transportation Research Board*, pages 1–16.

Dee, H. M. and Velastin, S. A. (2007). How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, 19(5-6):329–343.

Fatah, A. and Mottram, C. (2007). Collective Choreography of Space: Modelling Digital Co- presence in a Public Arena. In *1st Symposium on Systems Research in the Arts and Humanities "On Choreographies in Music, Visual and Performing Arts, and Environmental Design"*, pages 59–63, Baden-Baden.

Feldmann, S., Kyamakya, K., Zapater, A., and Lue, Z. (2003). An indoor Bluetooth-based positioning system: concept, implementation and experimental evaluation. In *International Conference on Wireless Networks (ICWN)*, pages 109–113, Las Vegas, NV.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., and Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160:60–68.

Hay, S. and Harle, R. (2009). Bluetooth Tracking without Discoverability. In *4th International Symposium on Location and Context Awareness*, pages 120–137, Tokyo, Japan.

Helbing, D., Johansson, A., and Al-Abideen, H. (2007). Dynamics of crowd disasters: An empirical study. *Physical Review E*, 75(4).

Hossain, A. and Soh, W. (2007). A comprehensive study of Bluetooth signal parameters for localization. In *18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, Athens, Greece.

Hui, S. K., Fader, P. S., and Bradlow, E. T. (2009). Path Data in Marketing : An Integrative Framework and Prospectus for Model Building. *Marketing Science*, 28(2):320–335.

Jansen-Verbeke, M., Van Rompaey, V., De Greve, K., and Verhamme, T. (2003). Gentse Feesten: effectenmeting als beleidsinstrument. Technical report, Steunpunt Toerisme en Recreatie, Leuven.

Kelly, D. (2010). *Minimal Infrastructure Radio Frequency Home Localisation Systems.* Phd thesis, National University of Ireland Maynooth.

Kemperman, A., Borgers, A., and Timmermans, H. (2009). Tourist shopping behavior in a historic downtown area. *Tourism Management*, 30(2):208–218.

Kostakos, V. and O'Neill, E. (2008). Capturing and visualising Bluetooth encounters. In *Adjunct proceedings of the Conference on Human Factors in Computing Systems (CHI 2008)*, Florence.

Laube, P., Dennis, T., Forer, P., and Walker, M. (2007). Movement beyond the snapshot – Dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, 31(5):481–501.

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Leitinger, S., Gröchenig, S., Pavelka, S., and Wimmer, M. (2010). Erfassung von Personenströmen mit der Bluetooth-Tracking- Technologie. In *Angewandte Geoinformatik 2010*, pages 220–225, Salzburg, Austria.

Millonig, A. and Gartner, G. (2008). Shadowing – Tracking – Interviewing: How to Explore Human Spatio-Temporal Behaviour Patterns. In Gottfried, B. and Aghajan, H., editors, *Proceedings of the 2nd Workshop on Behaviour Monitoring and Interpretation (BMI '08)*, volume 396, pages 1–14, Kaiserslautern.

O'Connor, A., Zerger, A., and Itami, B. (2005). Geo-temporal tracking and analysis of tourist movement. *Mathematics and Computers in Simulation*, 69(1-2):135–150.

Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. (2007). Trajectory data warehouses: Design and implementation issues. *Journal of Computing Science and Engineering*, 1(2):240–261.

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A., Penn, A., Fraser, D., and Jones, T. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *8th International Conference on Ubiquitous Computing (UBICOMP 2006)*, pages 315–332, Orange County, CA.

Peterson, B., Baldwin, R., and Kharoufeh, J. (2006). Bluetooth inquiry time characterization and selection. *IEEE Transactions on Mobile Computing*, 5(9):1173–1187.

Quinlan, E. (2008). Conspicuous Invisibility: Shadowing as a Data Collection Strategy. *Qualitative Inquiry*, 14(8):1480–1499.

Quiroga, C. A. and Bullock, D. M. (1998). Travel time studies with global positioning and geographic information systems: an integrated methodology. *Transportation Research Part C: Emerging Technologies*, 6(1-2):101–127.

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

Rice, R. E. and Katz, J. E. (2003). Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts. *Telecommunications Policy*, 27(8-9):597–623.

Saxena, S., Brémond, F., Thonnat, M., and Ma, R. (2008). Crowd behavior recognition for video surveillance. In *Advanced Concepts for Intelligent Vision Systems (ACIVS) Conference*, pages 970–981, Juan-les-Pins.

Shaw, S.-L., Yu, H., and Bombom, L. S. (2008). A Space-Time GIS Approach to Exploring Large Individual-based Spatiotemporal Datasets. *Transactions in GIS*, 12(4):425–441.

Shoval, N. and Isaacson, M. (2007a). Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*, 97(2):282–297.

Shoval, N. and Isaacson, M. (2007b). Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1):141–159.

Terry, J. P. and Feng, C.-C. (2010). On quantifying the sinuosity of typhoon tracks in the western North Pacific basin. *Applied Geography*, 30(4):678–686.

Van Londersele, B., Delafontaine, M., and Van de Weghe, N. (2009). Bluetooth Tracking. *GIM International*, 23(11):23–25.

Van Schaick, J. and Van der Spek, S., editors (2008). *Urbanism on Track - Application of tracking technologies in urbanism.* IOS Press, Amsterdam.

Versichele, M., Delafontaine, M., Neutens, T., and Van de Weghe, N. (2010). Potential and Implications of Bluetooth Proximity-Based Tracking in Moving Object Research. In *1st International Workshop on Movement Pattern Analysis (MPA) in conjunction with the 6th International Conference on Geographic Information Science (GIScience)*, pages 111–116, Zurich, Switzerland.

Wasson, J. S., Sturdevant, J. R., and Bullock, D. M. (2008). Real-Time Travel Time Estimates Using Media Access Control Address Matching. *ITE Journal (Institute of Transportation Engineers)*, 78(6):20–23.

Weinzerl, J. and Hagemann, W. (2007). Automatische Erfassung von Umsteigern per Bluetooth-Technologie. *Nahverkehrspraxis*, 3:18–19.

Wilson, W. C. (1998). Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, 30(6):1017–1038.

Zeitz, K. M., Tan, H. M., Grief, M., Couns, P. C., and Zeitz, C. J. (2009). Crowd behavior at mass gatherings: a literature review. *Prehospital and Disaster Medicine*, 24(1):32–38.

Zhou, S. and Pollard, J. (2006). Position Measurement using Bluetooth. *IEEE Transactions on Consumer Electronics*, 52(2):555–558.

3

Mobile mapping of sporting event spectators using Bluetooth sensors: 'Tour of Flanders 2011'

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., Van de Weghe, N.

***Sensors,* 12(10):14196-14213**

Abstract *Accurate spatiotemporal information on crowds is a necessity for a better management in general and for the mitigation of potential security risks. The large numbers of individuals involved and their mobility, however, make generation of this information non-trivial. This chapter proposes a novel methodology to estimate and map crowd sizes using mobile Bluetooth sensors and examines to what extent this methodology represents a valuable alternative to existing methods. The proposed methodology is applied in a unique case study that uses Bluetooth technology for the mobile mapping of spectators of the 'Tour of Flanders 2011' road cycling race. The locations of nearly 16,000 cell phones of spectators along the race course were registered and detailed views of the spatiotemporal distribution of the crowd were generated. Comparison with visual head counts from camera footage delivered a detection ratio of 13.0±2.3%, making it possible to estimate the crowd size. To our knowledge, this is the first*

study that uses mobile Bluetooth sensors to count and map a crowd over space and time.

3.1 Introduction

Throughout human history, people have had the tendency to sometimes gather in large numbers — either in an organized or spontaneous way. The character of these gatherings or the motivations of the people present may vary from purely recreational (festivals, parades, sports), religious (pilgrimages) to political (demonstrations, inaugurations) as described by Getz (2008). Regardless of the expected level of agitation, large crowds will always constitute a potential security risk. Indeed, they will sometimes reach or even surpass the short-term carrying capacity of the local environment or certain bottlenecks inside it. Often caused by lacking planning and management, critical crowd densities can give rise to significant human casualties (Zhen et al., 2008). Besides these security issues, large events additionally provide significant economical opportunities (Kasimati, 2003; Prentice and Andersen, 2003). Certain mega events such as the Olympics also have important social impacts for the hosts (Waitt, 2003).

Whether for crowd management or economic incentives, one of the most important (and certainly most tangible) indicators of a crowd is its size. When access to an event is restricted (e.g. through tickets or turnstiles at access points), *counting* a crowd is trivial. For open-access events, however, this is much more challenging. Additionally, crowd size estimations often differ from reality due to subjectivity or contradicting motives of the different stakeholders (Watson and Yip, 2011). Given that the success of an organized event (e.g., a protest march) is often measured by its attendance, organizers may be tempted to exaggerate attendance figures in order to put more weight on public opinion. Perhaps one of the most telling examples is that of the 'Million Man March' held in 1995 in Washington DC where, depending on the source, crowd size estimations varied between 400,000 and 1.5—2 million.

Public safety cannot afford such margins of error and requires objective and accurate information. While various methodologies (an overview is given in section 3.2 on the next page) have been suggested to estimate the size of a crowd in an objective manner, they often entail high levels of uncertainty and are impractical when applied in scenarios with high levels of movement. The *mapping* of a crowd, for which information is needed on the specific location or sequence of locations of individuals within the crowd, is even more complex than just *counting* and requires more advanced methodologies.

In this chapter, we will present an alternative methodology for counting and mapping a crowd based on the Bluetooth technology. The usefulness of our approach will be illustrated in a case study where spectators of a road cycling race are mapped using Bluetooth sensors installed on a mobile platform moving along the track, delivering detailed spatiotemporal information on the crowd assembled for the sporting event. In section 3.2 on the facing page, we discuss the current methodologies used to count and/or map crowds, their most important deficiencies and how Bluetooth technology might offer a valuable alternative — especially when individual mobility needs to be accounted for. Subsequently, we present the Bluetooth tracking methodology and its specific deployment in section 3.3 on page 50. In section 3.4 on page 51, we then give background information on the case study ('Tour of Flanders 2011'). The results of an experiment carried out prior to the cycling race, and the main case study experiment itself are outlined in section 3.5 on page 53. Finally, we interpret and discuss these results (section 3.6 on page 60) and give a short conclusion (section 3.7 on page 62).

3.2 Counting and mapping a crowd

Different methodologies have been proposed to estimate crowd sizes. First, rough estimations can be made by speculating on the basis of prior experience and knowledge of the local terrain, or manually counting either static or mobile attendees at one or more fixed locations (Yip et al., 2010). Alternatively, secondary data sources such as the amount of waste generated by a public and public transport usage to reach an event (Versichele et al., 2012b) have also been used in the absence of readily available primary data. A third and more sophisticated methodology — introduced in the sixties (Jacobs, 1967) and later modified in the seventies (Seidler et al., 1976) — is to carefully analyze aerial photographs of a crowd and to outline zones of uniform crowd density. Using standard density rules that are still used today (loose crowd: 1 $person/m^2$, solid crowd: 2 $persons/m^2$, very dense crowd: 4 $persons/m^2$) and the surface areas of the outlined zones, one can estimate the total number of attendees. For the previously mentioned 'Million Man March'[1], this grid/density methodology yielded an estimate of 870,000 people with a margin of error of about 25%. Several other studies have finally calculated crowd densities with the help of computer vision techniques on very high resolution satellite images (Sirmacek and Reinartz, 2011) or ground-based cameras (Kong and Gray, 2005; Rahmalan et al., 2006). Despite some promising results, these techniques remain confined to laboratory conditions (Dee and Velastin, 2007). Hence, there is a need for a more robust methodology.

Counting a crowd gets even more challenging, when the dynamics of the crowd are to be accounted for. In the relevant literature, mobility is usually attributed to the crowd itself (e.g. a march), giving rise to a distinction between static and mobile crowds, with different counting methodologies for each of these categories (Watson and Yip, 2011). Mobility can, however, also be part of a scenario with a (largely) static crowd when there is a mobile 'attractor' at play (e.g. a parade or a cycling race where spectators are lined up along a linear trajectory). As such, both the mobility of the crowd and the attractor (if present) should be taken into account. Table 3.1 on the following page summarizes how different crowd scenarios may be formed based on the above distinctions.

[1] http://www.bu.edu/ remotesensing/research/completed/million-man-march/

Table 3.1: Characterization of crowd scenarios according to the mobility of the attendees and the presence/mobility of an attractor. The attractors for the specific examples are shown between brackets.

		Attractor		
		Static	Dynamic	No attractor present
Attendees	Static	inauguration (president), football game (pitch)	cycling race (cyclists), papal visit (popemobile)	New Years' celebration
	Mobile	riot (police forces)	love parade (trucks with music installations)	marathon, demonstration

The added difficulty in estimating the size of a dynamic crowd has previously been studied. In a demonstration, for example, manual head counts at fixed locations were found to be labor-intensive, error prone and cannot account for people leaving a march in front of or entering a march behind a counting location (Yip et al., 2010). Even if there are good photographs of a mobile crowd available for a grid/density estimation, the area occupied by a dynamic crowd is difficult to define (Watson and Yip, 2011). All of the above-mentioned methodologies have the additional drawback that they only generate a snapshot view of the crowd size, ignoring its dynamic nature.

As it appears, said methodologies have significant limitations in terms of counting crowds, and are ill-suited to map crowds onto space and/or time due to their single snapshot view. Recent technologies able to register individual movements have the capacity to fill this methodological gap. Several of these tracking technologies in different developmental stages have been used to date. Although computer vision techniques have already been able to reconstruct individual trajectories inside a crowd (Marrón-Romera et al., 2010; Rabaud and Belongie, 2006), correctly applying these techniques in real-life and large-scale scenarios is still beyond the current state of the art due to several reasons including the necessity of a multitude of camera views, occlusions, variable weather conditions, etc. (Dee and Velastin, 2007). Other methodologies take advantage of the growing adoption of positioning technologies on modern smartphones, with the most prominent example being the global positioning system (GPS). While this technology is able to deliver fine-grained movement data of individuals, the need for active cooperation of the traced individual — either by installing an application on the smartphone or by distributing logging devices (Van der Spek et al., 2009) — makes it labor-intensive and less feasible when a representative sample of a large crowd is to be studied. The movement of a mobile device can also be reconstructed by using log files of mobile operators containing information about which cell-towers the device connected to during its lifetime or when calls were made (Ahas et al., 2008; González et al., 2008). The locational precision of this last methodology (in the order of at least 100 meters even in urban settings) is, however, too large for handling the smaller-scale dynamics of crowds (Ahas et al., 2007).

More recently, Bluetooth has been suggested as an interesting alternative tracking technology. Since the Bluetooth protocol allows for wireless discovery and identification of nearby devices, static Bluetooth sensors placed at strategic locations can give insights into human mobility in a variety of contexts: dynamics at mass events (Stange et al., 2011; Versichele et al., 2012a), urban design (O'Neill et al., 2006), social studies (Kostakos and O'Neill, 2008), travel time estimation of motorized traffic (Haghani et al., 2009), etc. Initially envisioned as a low-power and open protocol for implementing Wireless Personal Area Networks by Siemens in 1994, Bluetooth has since become an almost ubiquitous technology on modern mobile devices. Prior to the ability for two devices to connect wirelessly through Bluetooth, one device needs to be discovered by the other. This part of the Bluetooth protocol is called the 'inquiry phase' (Peterson et al., 2006). The master device transmits inquiry packets, to which discoverable devices within its vicinity respond with inquiry response packets. These include the MAC address (which is a 48-bit identifier of the mobile device), and the class of device (COD) code (which gives a general idea about the type of device and some of its functionalities). By mapping detected MAC addresses to a specific timestamp and location where a sensor that made the discovery was located, one can reconstruct proximity-based trajectories (Bensky, 2007). Since an actual connection is not required, tracked individuals are not aware of the presence of Bluetooth sensors and the methodology is in essence completely unobtrusive. Since Bluetooth 1.2, it is also possible to register the received signal strength indicator (RSSI) of the inquiry response packets, which is loosely correlated with the distance between the sensor and the detected device (Hossain and Soh, 2007).

In this chapter, we propose a novel use-case for the Bluetooth tracking methodology: the mapping of spectators along the track of a road cycling race. We build on the concept of 'mobile mapping', where a combination of a moving platform, navigation sensors and mapping sensors is used for the geo-referenced mapping of information across a study area (Li, 1997). Instead of the more traditional imaging sensors, however, we use Bluetooth sensors to detect the proximity of people carrying Bluetooth-enabled mobile phones. The methodology bears resemblance to the popular act of 'war driving' where the locations of WiFi Access Points are mapped by a car driving around a study area (Berghel, 2004). The idea of using mobile Bluetooth sensors is not novel as such. Particularly social studies have already embraced the concept of mobile phones as wearable (Bluetooth) sensors for investigating complex social systems (Eagle and Pentland, 2005) or 'familiar strangers' as individuals we repeatedly observe yet do not directly interact with (Paulos and Goodman, 2004) . Other studies have already hinted at the possible use for discovering pedestrian travel behavior as well (Malinovskiy et al., 2012). To our knowledge, however, the concept of using a mobile (Bluetooth) sensor to map a crowd along a trajectory followed by a mobile attractor is without a precedent in scientific literature.

3.3 Methodology and deployment

In order to map spectators, we used a mobile platform equipped with two Bluetooth sensors that moved along the track registering Bluetooth devices belonging to spectators as it passed them by. Figure 3.1 on the next page shows a conceptual representation of the methodology at the top and the used equipment at the bottom. The numbers in the figure correspond to the numbers between brackets in this paragraph. The mobile platform carrying the equipment was a car (Kia Sportage) that belonged to the convoy preceding the racers (on average the racers lagged the platform for between 3 and 6 minutes). The Bluetooth sensors (SENA Parani UD-100) were attached to the side windows in the back of the car (1). These class 1 Bluetooth devices (i.e. the most powerful class with a theoretical communication range of 100 meters) were fitted with an external stub antenna with a gain of 1 dB_i. Previous experiments had shown that this combination of sensor and antenna is capable of discovering mobile phones at distances of 100 meters in a static context. The Bluetooth sensors performed new inquiry scans every 10.24 seconds. Both sensors were connected to a portable computer (Dell Vostro 3500, *Debian Testing OS*, kernel 2.6.38-2) running Bluetooth scanning software called *Gyrid*[2] (version 0.4.5). This self-implemented framework, built around the *PyBlueZ* (0.18-1) and *BlueZ* (version 4.89-1) frameworks, was also used in previous studies of our research group. Additionally, a video camera (Panasonic HDC-SD10) was installed looking through the front window (2). This way, we could later compare Bluetooth counts with visual counts. Because the recorded video in raw format was too large to fit onto one memory card, two memory cards were regularly swapped and their content copied. Finally, a GPS unit (Garmin GPS60) was used to continuously record the position of the mobile platform (3). A logger on the portable computer registered these locations on the fly.

Every time a Bluetooth device was detected, its MAC address, COD code, the RSSI of the detection and the timestamp of the detection were registered. Because of the GPS recordings, every detection at a certain timestamp could later be mapped onto a location along the track followed by the car. Because both the GPS unit as well as the Bluetooth sensors were connected to the same computer, their outputs were automatically synchronized in time. The computer itself was connected to the Internet through a 3G connection and was synchronized with the Network Time Protocol. Synchronization between video data and Bluetooth data was done by manual landmark discovery in the camera footage and linking this to their known position and the time they were passed according to the GPS data. As is indicated in figure 3.1 on the facing page by some of the spectators being mapped onto several nearby positions along the track, some devices are detected more than once during a passing event. Terminologically speaking, these devices are associated with several detections. This important distinction leads to the concept of a set of (unique) *devices* and an associated larger set of (non-unique) *detections*. Additionally, it is known that a minority of spectators will watch the group of cyclists at more than one location along the track. This

[2] https://github.com/Roel/Gyrid

is depicted by the arrow in the figure. Finally, only a subset of the spectators owns a mobile device with a visible Bluetooth interface and the rest of the crowd is hence not detected (indicated by the transparent icons).

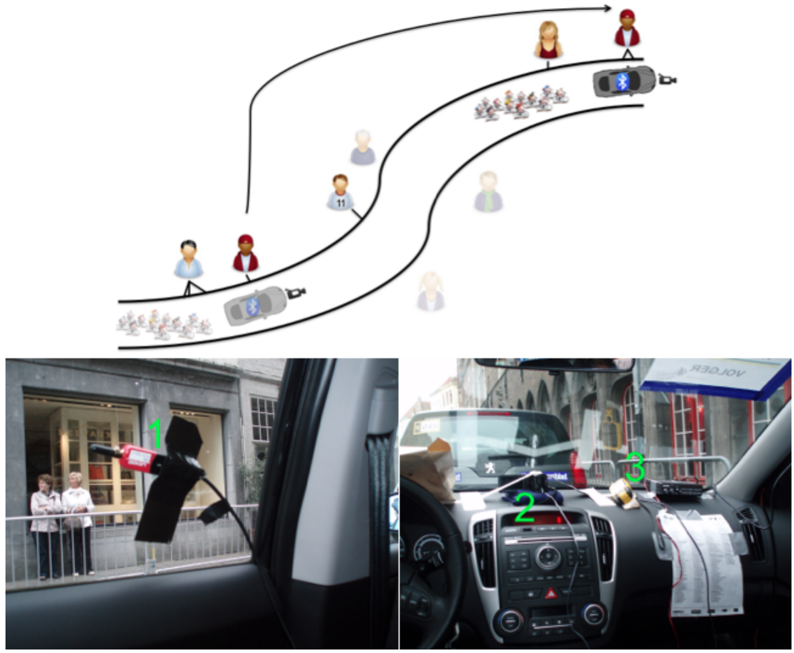


Figure 3.1: Mobile sensor deployment during the race. Top: schematic overview of the methodology, bottom: actual deployment (1: Bluetooth sensor, 2: video camera, 3: GPS unit).

3.4 Case study: 'Tour of Flanders 2011'

The 'Tour of Flanders' (in Dutch: 'Ronde van Vlaanderen') is a one day road cycling race held yearly in Flanders, Belgium. Due to the popularity of cycling in Belgium and the long history and tradition of the race (the first edition dates back to 1913), the event has become the largest sporting event in Flanders and, as such, has become more or less part of the local cultural heritage. Due to the open nature of the event (the race track only traverses public space) little is known about the spatiotemporal characteristics of spectators. Correspondingly, the race was chosen as a fitting case study for the application of Bluetooth tracking in a mobile mapping context.

An overview of the trajectory followed by the mobile platform — including its speed — is given in figure 3.2 on the next page. The 2011 edition started in Bruges at 9:45 am and ended in Meerbeke (winner: 16:00 am, last rider: 16:18 am), taking the cyclists over 256.3 kilometers and 18 slopes. For safety reasons, some sections of the race could not be accessed by fourwheeled vehicles. Therefore, the trajectory of the mobile platform carrying the Bluetooth sensors in some places deviates from the official track. This was the case around four slopes (8, 9, 13 and 17). As a consequence, the crowds gathered at these locations were not covered by the platform. There were two points along the track where racers were provided food and drinks.

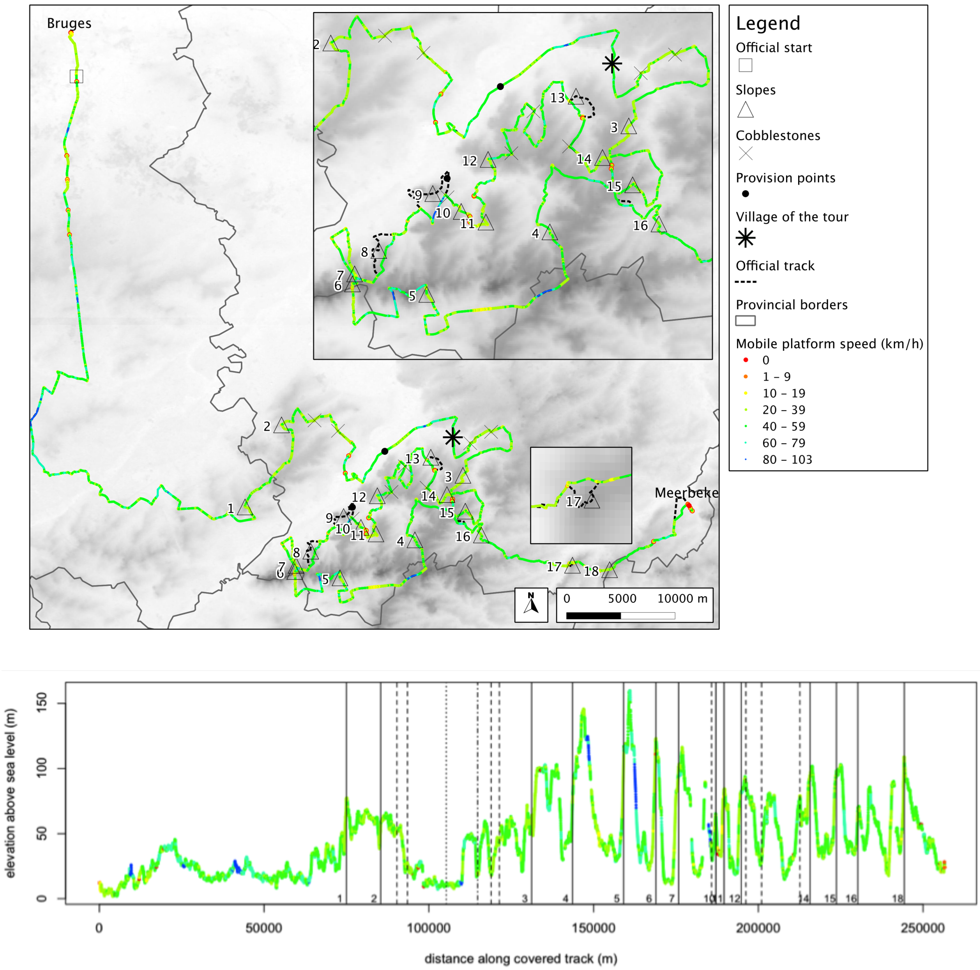


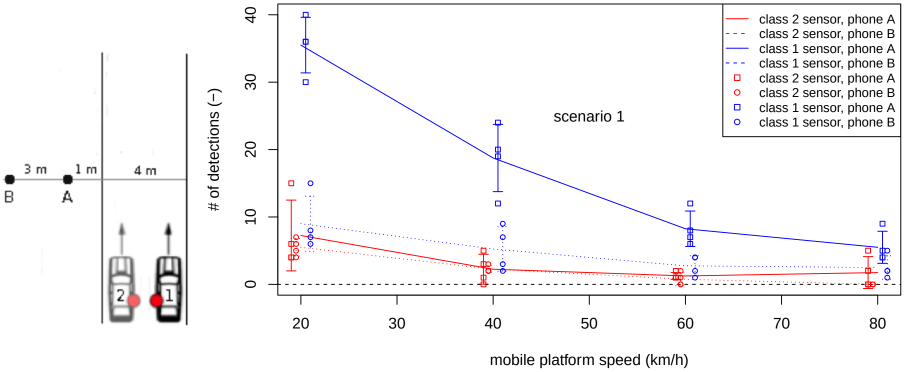
Figure 3.2: 'Tour of Flanders', 2011 edition. Top: spatial view of the official track and the trajectory of the mobile platform. Bottom: elevation profile of the mobile platform trajectory (NASA Shuttle Radar Topography Mission elevation data). In both cases, the speed of the mobile platform is shown in a color scale ranging from red (temporary stops) to blue (more than 80 km/h). The 18 official slopes are shown as triangles in the spatial view and solid vertical lines in the elevation profile (not covered slopes excluded). The dashed vertical lines represent the cobblestoned segments, the dotted vertical line the first provision point, and the dashed-dotted line the village of the tour.

3.5 Results

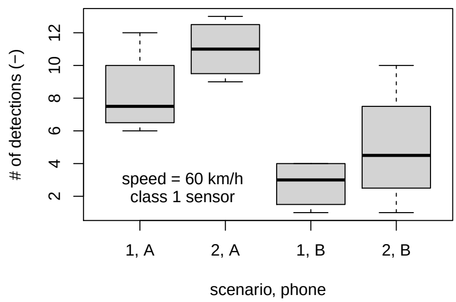
3.5.1 Prior experiments with mobile platform

Prior to deploying the mobile platform, the Bluetooth detection process and its sensitivity towards several factors in a mobile context needed to be investigated. Five factors were studied: (1) the type of Bluetooth sensor, (2) its position on the mobile platform, (3) the distance between the detectable mobile phone and the road covered by the mobile platform, and (4) the speed of the mobile platform. A small-scale experiment was conducted under controlled conditions by placing two discoverable Bluetooth-enabled phones (representing two spectators) next to a road section at different distances (resp. 1 and 3 meters perpendicular distance, at 1 meter above ground), controlling the different factors and calculating the number of detections of each phone during a passing event of the mobile platform. For every combination of factors, four passage runs were made. The main goal of this experiment was to check the feasibility of the proposed methodology before deploying it on a larger scale. A view of the experimental setup is given in figure 3.3a on the following page. The influence of the sensor type, the mobile platform speed and the distance between the road and the mobile phone can be seen in figure 3.3b on the next page. The difference in performance between the class 2 and class 1 sensor is evident. Where the class 2 sensor already missed the closest phone once at 40 km/h (the furthest phone was even missed on every run at 80 km/h), the class 1 sensor did not miss any phone on any run. The mobile platform speed influences the number of detections in a negative way. More importantly, though, it does not seem to form a bottleneck when using a class 1 sensor (at least for speeds not surpassing 80 km/h). The difference between both phones is manifest in the case of the class 1 sensor: a larger distance lowers the number of detections. With a class 2 sensor, the difference is nearly negligible.

Next, the position of the sensor on the mobile platform needed to be investigated. Since it was not possible to mount an antenna on the rooftop of the car, it had to be placed inside of the car. Prior experiments had shown the rear side window being the best location in contrast to the ceiling or the front windshield (not shown). It was not known, however, what the effect was of the sensor either facing or not facing the side of the road with detectable phones. This was tested by performing passage runs at a speed of 60 km/h under both scenarios (figure 3.3c on the following page). Contrary to what could be expected, both phones were detected more often when the sensor was not facing them (scenario 2) although the difference is clearly not statistically significant. The difference between both phones is more pronounced, showing that the effect of the distance is larger than that of the sensor placement in the car. In order to minimize the probability of missing detectable phones, the mobile platform was equipped with a class 1 sensor on each side during the actual case study experiment described in the next section.



(a) **(b)**



(c)

Figure 3.3: Mobile Bluetooth detection process investigation: (a) experimental setup showing both mobile phones (A, B) and the two passing scenarios (1: sensor facing phones, 2: sensor not facing phones); (b) effect of sensor type, speed, and phone distance (both the mean and standard deviations, as well as the individual values of all combinations are shown; both error bars and individual data points are offset from their real x-value for visual clarity); (c) effect of sensor placement on mobile platform at a speed of 60 km/h.

3.5.2 Case study

3.5.2.1 Preprocessing and mapping of detections along trajectory

During the course of the race, two types of log files were generated. First, each Bluetooth sensor generated a Bluetooth log file containing one log line per detection in the following format: *timestamp, MAC address, COD code, RSSI* (e.g. '20110403-101520, 00:12:34:56:78:9A, 5898756, -72'). Second, the sequential GPS fixes were gathered on the portable computer and written to a separate log file. In order to geo-localize the Bluetooth detections, both data sources needed to be combined. First of all, the Bluetooth detections picked up before the GPS unit started recording were deleted. Subsequently, the data were imported in the *R* suite (version 2.14.0) and transformed to two data frames (one for the Bluetooth data of both sensors merged together and another for the GPS data). Table 3.2 lists some of the main properties of both data frames. Given that the GPS unit had a sampling rate of 1 second and the Bluetooth detections had a temporal precision of 1 second as well, both data frames could be merged into one data frame based on their respective timestamps. As such, each Bluetooth detection was given a location (both an xy-location and a distance along the trajectory followed by the mobile platform). Finally, we examined the device classes detected by the sensors and found that roughly three out of four detections were from phones, and one out of four were audio/video devices. Other types (including devices which did not give COD information) were very sparse. Only phones can be more or less directly linked to a physical person. So in order to map only spectators, we filtered out the 15,597 phones (which amounted to 130,464 detections in the dataset). The set of audio/video devices consisted almost entirely of handsfree devices such as car kits and represent vehicles rather than persons. They were excluded from the rest of the analyses together with the other less frequent device classes.

Table 3.2: Data preprocessing summary.

Bluetooth	First detection	09:34:06	
	Last detection	16:13:57	
	# detected devices	16,182	
	# detections	177,079	
GPS	First fix	09:34:06	
	Last fix	16:14:46	
	Total distance	256,668 m	
Video	Start	09:40:51	
	End	15:52:50	
	Blind spots (swapping of memory cards)	$\pm$ 30 min	
Filtering	**Device class**	**# detections**	**# detected devices**
	Phone	130,464 (73.7%)	15,597 (96.4%)
	Audio/Video	45,073 (25.5%)	302 (1.9%)
	Unknown	787 (0.4%)	124 (0.9%)
	Computer	598 (0.3%)	136 (0.8%)
	Network access point	122 (0.1%)	16 (0.1%)
	Imaging	35 (0.0%)	7 (0.0%)

3.5.2.2 Crowdedness along trajectory

In order to map spectators along the trajectory, we divided the trajectory into 1 kilometer long segments. For each segment, we aggregated the Bluetooth detections that occurred in the same time span and calculated the number of unique MAC addresses. This way we could map the number of detectable spectators along the course of the trajectory as an indicator of local crowdedness. The result is shown in figure 3.4 on the facing page. The two-dimensional spatial view and one-dimensional cross section show alternating zones of higher and lower numbers of spectators. Most of the densely crowded segments coincide with the slopes along the track. Slope 7 ('Oude Kwaremont') clearly attracted the largest number of detectable spectators (582 phones over 1 kilometer). The other crowded segments are either associated with cobblestoned segments, or villages or cities (visible as concentrations with high population densities) that are located on the track.

3.5.2.3 From Bluetooth devices to crowd size

In the previous section, we have identified crowded zones according to the number of phones detected over a certain distance. As stated in the introduction, however, we intend to make reasonable estimations of the size of an entire crowd. As a consequence, we have to transform the number of detected devices into an approximate number of spectators. In order to do this, we need to know how large the share of detectable persons in a crowd is. We call this the 'detection ratio', and calculate it by counting the number of (unique) detected devices and visual head counts during a certain time period, and dividing the former by the latter. In case of static Bluetooth sensors, a person usually sits in the close vicinity of a sensor and counts the number of people passing by. Using this methodology, a previous study with a dataset gathered in 2010 delivered a detection ratio of 11.0±1.8% (Versichele et al., 2012a). Visually counting people on the side of a road from a mobile platform is less trivial than counting moving people from a static position. Nevertheless, the video footage recorded from the mobile platform could be used to count spectators visible from the platform as it passed by. In total, 14 calculations were made along the trajectory covering 52.2 kilometers in total (20% of the total trajectory length). These are shown in table 3.3 on page 58. If all measurements are taken into account, the average detection ratio lies at 14.3±3.9%. The resulting relative standard error RSE, which represents the relative error the crowd size estimation will have, is 27.3%. If the two measurements with abnormally high detection ratios (measurements 6 and 9 both lie above $Q3$ (the third quartile) + IQR (interquartile range)) are regarded as outliers and excluded, we end up with an average detection ratio of 13.0±2.3% (RSE of 17.9%). Further clarification as to why these two values can be considered outliers is given in the discussion. The table also includes the detection ratio of a previous experiment with static Bluetooth sensors (Versichele et al. (2012a), RSE of 16.4%). Using these detection ratios, the extrapolated total crowd size at the segment with the largest number of detected phones along the trajectory lies at 4,070 ([3,198–5,596], with outliers) or 4,477 ([3,804–5,439], without outliers). We can use these figures to roughly get an idea about the spectator density along the road. Taking into account the length of this segment (1,000 meters) and the fact that spectators can line up on both sides of the road, we end up with roughly 2 spectators per meter of roadside. Analogously, the estimate of the total crowd size

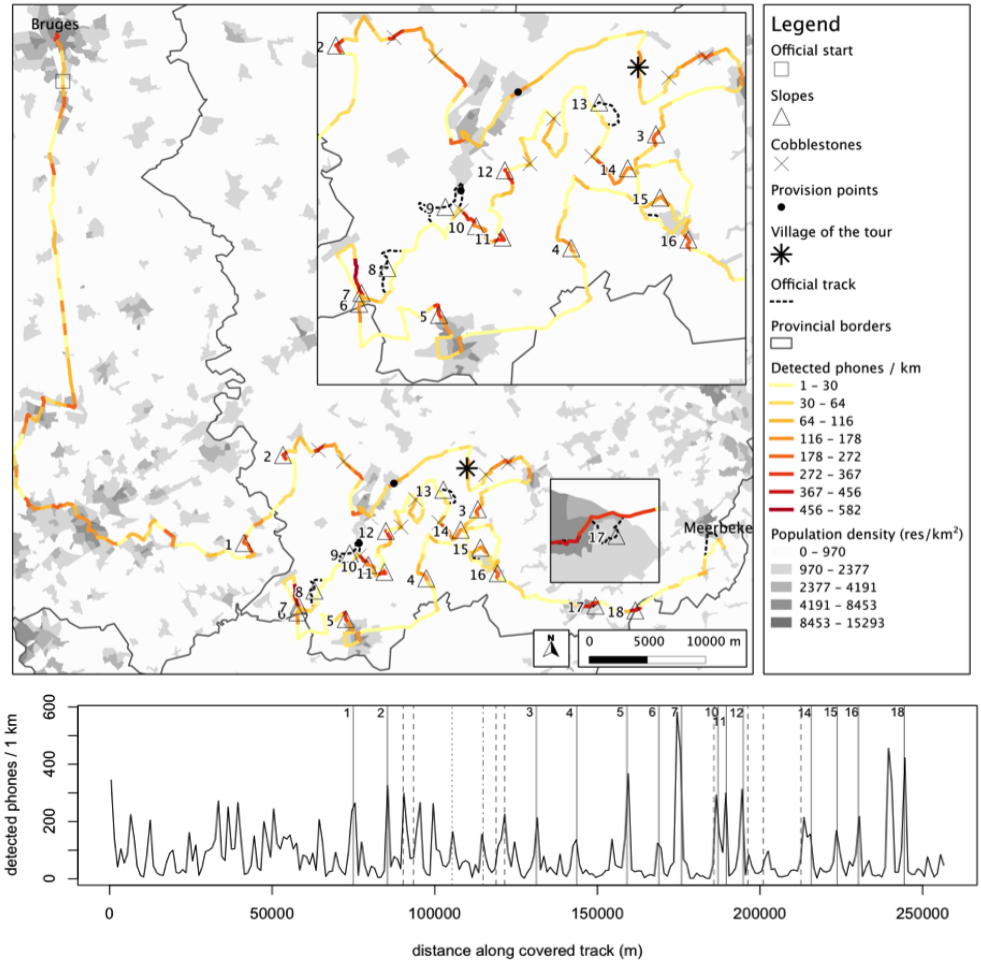


Figure 3.4: Number of detected phones along 1 kilometer long segments of the trajectory followed by the mobile platform as an indicator of crowdedness. Top: spatial view using a yellow-to-red color scale to depict the number of phones, class breaks according to the Jenks natural breaks (Jenks and Caspall, 1971) optimization. The background shows the population density of statistical sectors in the area, colored in a grey-scale, once again according to Jenks natural breaks. Bottom: one-dimensional view.

that was covered by the mobile platform during the entire duration of the measurements lies at 109,070 ([85,720–149,957], with outliers) or 119,977 ([101,549–145,937], without outliers). This represents roughly one spectator every 2 meters.

Table 3.3: Detection ratios along the trajectory, calculated by comparing the numbers of detected Bluetooth phones with visual spectator counts from video recordings. The two rows in *italics* represent abnormally high detection ratios and can be regarded as outliers. The bottom part of the table shows the average detection ratios, standard deviations and relative standard errors with and without pruning of the outliers.

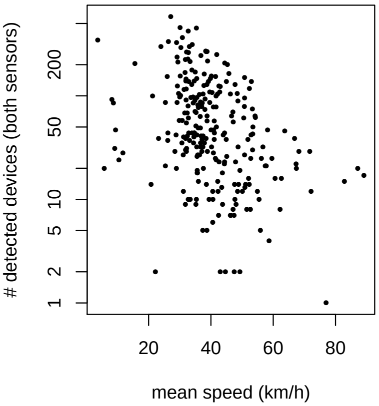
Number	**Duration** (s)	**Distance** (m)	**Visual count** (-)	**Detected phones** (-)	**Detection ratio** (%)
1	447	3,730	1,984	253	12.75
2	527	4,246	1,350	153	11.33
3	1,035	11,694	4,567	580	12.70
4	286	4,909	1,919	313	16.31
5	144	3,148	1,013	116	11.45
6	*143*	*1,729*	*1,034*	*220*	*21.28*
7	358	4,479	1,502	154	10.25
8	439	5,925	4,183	515	12.31
9	*275*	*2,777*	*1,207*	*273*	*22.62*
10	261	2,547	1,320	175	13.26
11	143	1,755	1,758	209	11.89
12	156	1,691	970	141	14.54
13	152	2,442	808	90	11.14
14	123	1,155	496	91	18.35

Average (%)	**St. dev.** (%)	**Rel. st. error** RSE (%)	
14.3	3.9	27.3	all measurements
13.0	2.3	17.9	measurements 6 and 9 pruned
11.0	*1.8*	*16.4*	*Versichele et al. (2012a)*

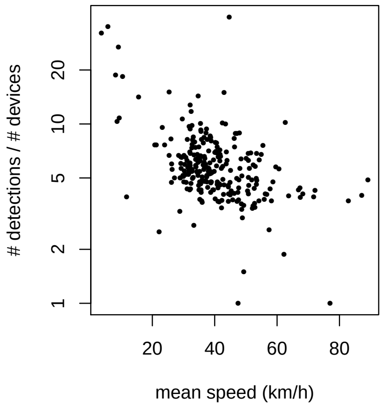
3.5.2.4 Mobile platform speed influence under real-life conditions

Since the Bluetooth sensors were placed on a mobile platform, we need to investigate whether the speed with which the platform moves along the trajectory has any influence on the spectator detection process. Preliminary experiments prior to the actual case study already demonstrated that the influence of the speed is negligible under idealized circumstances (section 3.5.1 on page 53). The dataset from the cycling race should, however, be explored as well in order to confirm or deny this finding. Four characteristics were calculated for each 1 kilometer long segment of the trajectory (as depicted in figure 3.4 on the previous page), and correlated with the mean speed attained over each segment. The resulting scatter plots are shown in figure 3.5 on the facing page.

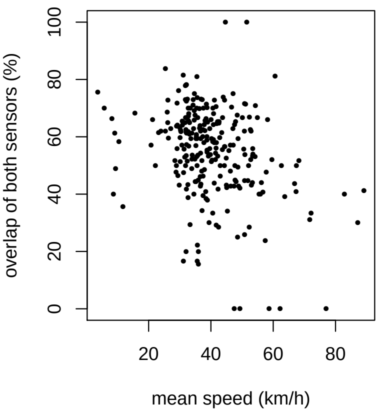
First, the number of detected phones over a segment is compared to the mean speed of the platform over that segment (figure 3.5a on the next page). The goal of this visualization is not to test for a direct link between both values (this assumption is not realistic anyway), but to focus on the segments at the low and high extremes of the speed spectrum. The segments with mean speeds over 80 km/h are still linked to between 10 and 20 detected phones, while the slowest segments are not associated with extremely high numbers of



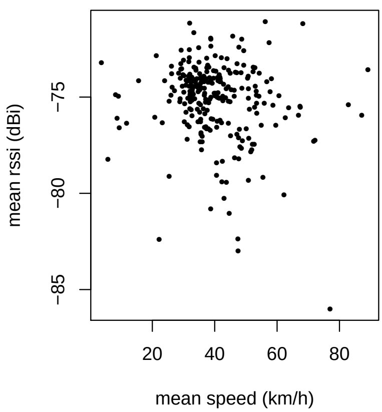
(a) $speed \sim \#detectedphones$



(b) $speed \sim \frac{\#detections}{\#detectedphones}$



(c) $speed \sim overlap$



(d) $speed \sim \overline{rssi}$

Figure 3.5: Real-life influence of mobile platform speed on four trajectory segment characteristics: (a) number of detected phones (*log scale*), (b) number of detections/number of detected phones (*log scale*), (c) sensor overlap, (d) mean RSSI.

detected phones. Consequently, we did not find an indication that very low or very high speeds bias the detection process in a significant way. Subsequently, the ratio between the number of detections and the number of detected phones is compared to the mean speed of each segment (figure 3.5b). Generally speaking, a higher ratio corresponds to phones being detected more times during a passing event (a higher sampling frequency) leading to a smaller likelihood of it being missed by the Bluetooth sensor passing it by. The very slow segments seem to correspond to slightly higher ratios, but the very fast segments have ratios that lie within the range of ratios for average speeds. This constitutes an extra indication that high speeds do not hinder the detection process in a significant way. The overlap between both sensors on the mobile platform (defined as the ratio of the size of the intersection of the sets of detected phones by each sensor to the size of the union of both device sets) seems to roughly lie between 20% and 80% (figure 3.5c on the previous page). Neither the slowest nor the fastest segments exhibit overlaps that deviate consistently with the general average. Finally, the mean of all RSSI values gathered in each segment is also plotted against the mean speed (figure 3.5d on the preceding page). Again, there is no indication of the signal strength of detections being influenced by the speed of the platform carrying the sensors.

3.6 Discussion

A crowd of spectators watching a road cycling race was mapped by a mobile platform equipped with two Bluetooth sensors. The locations of nearly 16,000 detected phones carried by the spectators were used as an indicator of the crowdedness levels along the race trajectory. Figure 3.4 on page 57 shows that the methodology was able to distinguish crowded zones from zones with sparse numbers of spectators. Nearly all hotspots corresponded to either slopes, cobblestoned segments or — in a lesser degree — urbanized areas. After careful examination of this map, the research team present during the race and the race organizers agreed that the results were consistent with their own experience.

Clearly, this reliability analysis cannot be constricted to personal (and possibly subjective) experience. Consequently, the camera footage from aboard the mobile platform was used as an additional (non-quantitative) ground truth data source. As an additional advantage of this data source, segments of the trajectory can be subjected to visual head counts forming smaller but quantitative data sources. By comparing these visual head counts with the number of detected phones, a detection ratio was calculated along 14 of these segments. This resulted in an average detection ratio of 14.3$\pm$3.9% (RSE of 27.3%). Most of the segments exhibited a detection ratio slightly higher than the reference detection ratio from a previous study performed in a static context ($\pm$11%), while two segments showed unusually high detection ratios above 20%. There are two possible explanations for this. First, it is possible that some phones located along neighboring segments are included in the detection ratio calculation while the visual head counts are restricted more strictly to the 1 kilometer long segment. This may lead to a higher enumerator and thus a higher detection ratio. Since we anticipated this effect when selecting the segments for visual head counts, the selected segments did not exhibit large crowds at their borders and therefore the effect should not be significant. A second and probably more significant effect is due to invisible spectators. Spectators who are missed for the visual head counts (e.g. in buildings) can also lead to an overestimation of detection ratios. This effect is much harder to control with an appropriate choice of segments because visual confirmation with the camera footage is impossible. The very high detection ratios are almost certainly caused by this effect. Correspondingly, an average detection ratio of 13.0$\pm$2.3% (RSE of 17.9%) is attained after pruning of the two most extreme outliers.

As already pointed out in the previous paragraph, we should consider the fact that a 1,000 meter long segment may actually cover a larger detection area. In the worst case this could measure up to 1,200 meters if the Bluetooth sensors would have a 100 meter detection range in a mobile context as well. The overlap between each pair of consecutive segments (defined as the ratio of their intersection and union) was extracted from the dataset. More than 90% of the pairs have overlaps smaller than the worst-case overlap of 10% ($^{100m\,+\,100m}/_{1000m\,+\,1000m}$). While the overlap between segments should not be completely disregarded, the effect will be of minor importance on larger scales.

An additional concern affecting the reliability is the selectivity of the detection process towards spectators of the race instead of the more general public (the entire race takes place on the public domain). In the worst case, the mobile platform could be mapping the population density because of a disproportionally large detection range. As a first way of evaluating to what extent this is the case, we included the population density as a background layer in figure 3.4 on page 57. While the generally higher numbers of detected devices in the south-western part of the map view could be linked to the locally higher population densities, most of the crowded zones further along are situated in locations with very low statistical population densities. Figure 3.6 shows a scatter plot of population density, calculated for each segment by a spatial join operation in *ArcGIS* returning the mean if a segment crossed more than one population density area, versus the number of detected phones along each 1 kilometer long trajectory segment. There is no clear correlation between both variables, which is another indication that there is a high selectivity towards spectators of the race. In the end, however, the correct counting of a crowd also raises an ontological issue linked to a specific scenario: how do you define a spectator of the race (e.g. solely by his/her proximity to the race track or also by additional constraints?).

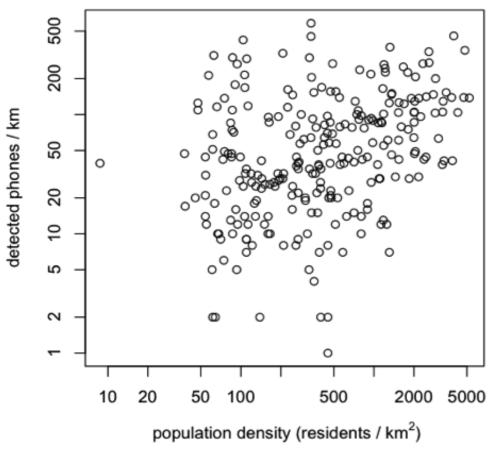


Figure 3.6: Scatter plot of statistical population density versus number of detected phones over each 1 kilometer long segment of the trajectory.

It is worthwhile comparing the Bluetooth tracking methodology with other and more often applied methodologies. Relevant literature (Watson and Yip, 2011) cites a relative standard error of the order of 10% for the grid/density method, seemingly making our methodology with a RSE of 17.9% less accurate. It should be stated, however, that it is first not always clear which ground-truth literature citations are based on. Additionally, the real merit of the proposed methodology is not just the counting of a crowd as such, but also its ability to easily map its behavior over time and space. As demonstrated in this study, Bluetooth tracking is able to generate insightful information (maps) of a crowd, its (varying) location, and size. With slight modifications, the same methodology could be used for crowds dispersed over more complex study areas. A mobile platform (or more than one) could also re-visit the same locations regularly for temporal pattern detection, or they could be combined with static sensors.

While the results were satisfactory in the context of this case study, other studies might necessitate more accurate counts and/or localizations of detections. Further calibration of the sensors will be necessary to accomplish this goal. The rather low overlap between both sensors on the platform, as seen in figure 3.5 on page 59, seems to indicate that a higher number of sensors/platforms might also play a beneficial role. Additionally, smaller scale investigations (in contrast with the larger-scale approach adopted in this case study) will need more certainty on the locational accuracy of the detections along the road section. The potential overlap between the detection areas of subsequent segments should also be further examined.

3.7 Conclusions and outlook

In this chapter, we demonstrated the added value of Bluetooth technology in the mobile mapping of individuals. The spectators of the 'Tour of Flanders 2011' were mapped along the trajectory followed by a mobile platform equipped with Bluetooth sensors. Nearly 16,000 devices were detected along 256 kilometers. Dividing the trajectory into 1 kilometer long segments, we were able to identify crowded hotspots in a detailed manner. These were mostly situated along slopes and cobblestoned segments, as anticipated by both the research group and the race organizers. Comparing the Bluetooth detection data with video recordings from the platform, an average detection ratio of 13.0$\pm$2.3% was calculated. The relative standard error of 17.9% is slightly higher than that of some alternative methodologies, but still acceptable for most purposes.

Our results showed that the main benefit of the proposed methodology is its ability to generate highly detailed spatiotemporal information on crowds with relative ease and without the need for expensive equipment or cooperation of the crowd. Despite these promising first results, some aspects were identified that need further investigation. A rather low overlap between the devices detected by both sensors on the mobile platform points out that a better calibration is necessary to further enhance the reliability of the detection process and its locational accuracy.

References

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Ahas, R., Laineste, J., Aasa, A., and Mark, U. (2007). The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In Gartner, G., Cartwright, W., and Peterson, M. P., editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 445–460. Springer, Berlin.

Bensky, A. (2007). *Wireless positioning technologies and applications.* Artech House, Boston, London.

Berghel, H. (2004). Wireless infidelity I: war driving. *Communications of the ACM*, 47(9):21–26.

Dee, H. M. and Velastin, S. A. (2007). How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, 19(5-6):329–343.

Eagle, N. and Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Getz, D. (2008). Event tourism: Definition, evolution, and research. *Tourism Management*, 29(3):403–428.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., and Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160:60–68.

Hossain, A. and Soh, W. (2007). A comprehensive study of Bluetooth signal parameters for localization. In *18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, Athens, Greece.

Jacobs, H. A. (1967). To count a crowd. *Columbia Journalism Review*, 6(1):37–40.

Jenks, G. F. and Caspall, F. C. (1971). Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61(2):217–244.

Kasimati, E. (2003). Economic aspects and the Summer Olympics: a review of related research. *International Journal of Tourism Research*, 5(6):433–444.

Kong, D. and Gray, D. (2005). Counting pedestrians in crowds using viewpoint invariant training. In *British Machine Vision Conference (BMVC 2005)*, Oxford, United Kingdom.

Kostakos, V. and O'Neill, E. (2008). Capturing and visualising Bluetooth encounters. In *Adjunct proceedings of the Conference on Human Factors in Computing Systems (CHI 2008)*, Florence.

Li, R. (1997). Mobile Mapping: An Emerging Technology for Spatial Data Acquisition. *Photogrammetric Engineering & Remote Sensing*, 63(9):1085–1092.

Malinovskiy, Y., Saunier, N., and Wang, Y. (2012). Pedestrian Travel Analysis Using Static Bluetooth Sensors. In *Transportation Research Board 91st Annual Meeting*, volume 250.

Marrón-Romera, M., García, J. C., Sotelo, M. A., Pizarro, D., Mazo, M., Cañas, J. M., Losada, C., and Marcos, A. (2010). Stereo vision tracking of multiple objects in complex indoor environments. *Sensors*, 10(10):8865–87.

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A., Penn, A., Fraser, D., and Jones, T. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *8th International Conference on Ubiquitous Computing (UBICOMP 2006)*, pages 315–332, Orange County, CA.

Paulos, E. and Goodman, E. (2004). The familiar stranger: anxiety, comfort, and play in public places. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, number 1, pages 223–230, Vienna, Austria.

Peterson, B., Baldwin, R., and Kharoufeh, J. (2006). Bluetooth inquiry time characterization and selection. *IEEE Transactions on Mobile Computing*, 5(9):1173–1187.

Prentice, R. and Andersen, V. (2003). Festival as creative destination. *Annals of Tourism Research*, 30(1):7–30.

Rabaud, V. and Belongie, S. (2006). Counting Crowded Moving Objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, volume 1, pages 705–711. IEEE.

Rahmalan, H., Nixon, M. S., and Carter, J. N. (2006). On crowd density estimation for surveillance. In *Institution of Engineering and Technology Conference on Crime and Security 2006*, pages 540–545, London.

Seidler, J., Meyer, K., and Gillivray, L. M. (1976). Collecting Data on Crowds and Rallies: A New Method of Stationary Sampling. *Social Forces*, 55(2):507–519.

Sirmacek, B. and Reinartz, P. (2011). Automatic crowd analysis from very high resolution satellite images. In *Photogrammetric Image Analysis Conference (PIA '11)*, pages 221–226, Munich.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Van der Spek, S., Van Schaick, J., De Bois, P., and De Haan, R. (2009). Sensing Human Activity: GPS Tracking. *Sensors*, 9(4):3033–3055.

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012a). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2):208–220.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., and Van de Weghe, N. (2012b). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10):14196–14213.

Waitt, G. (2003). Social impacts of the Sydney Olympics. *Annals of Tourism Research*, 30(1):194–215.

Watson, R. and Yip, P. (2011). How many were there when it mattered ? *Significance*, 8(3):104–107.

Yip, P. S. F., Watson, R., Chan, K. S., Lau, E., Chen, F., Xu, Y., Xi, L., Cheung, D. Y. T., Ip, B. Y. T., and Liu, D. (2010). Estimation of the Number of People in a Demonstration. *Australian & New Zealand Journal of Statistics*, 52(1):17–26.

Zhen, W., Mao, L., and Yuan, Z. (2008). Analysis of trample disaster and a case study – Mihong bridge fatality in China in 2004. *Safety Science*, 46(8):1255–1270.

4

Time-geographic derivation of feasible co-presence opportunities from network-constrained episodic movement data

Versichele, M., Neutens, T., Claeys Bouuaert, M., Van de Weghe, N.

Accepted for publication in *Transactions in Geography*

Abstract *Certain datasets on moving objects are episodic in nature — that is, the data is characterized by time gaps during which the position of the object is unknown. In this chapter, a model is developed to study the sparsely sampled network-constrained movement of several objects by calculating both potential and feasible (i.e. more likely) co-presence opportunities over time. The approach is applied to the context of a static sensor network, where the location of an object is only registered when passing a sensor location along a road network. Feasibility is incorporated based on the deviation from the shortest path. As an illustration, the model is applied to a large Bluetooth tracking dataset gathered at a mass event. The model output consists of maps showing the temporal evolution of the distribution of feasible co-presence opportunities of tracked visitors over the network (i.e. the number of visitors that could have been present together). We demonstrate the model's usefulness in studying the movement and distri-*

bution of a crowd over a study area with relatively few sampling locations. Finally, we discuss the results with a special emphasis on the distinction between feasible and actual presence, the need for further validation and calibration, and the performance of the implementation.

4.1 Introduction

Human or animal movement has traditionally been cumbersome to measure and register in an accurate way. Labor-intensive methodologies such as shadowing (Millonig and Gartner, 2008) and the distribution and recollection of travel diaries (Axhausen et al., 2002) are nowadays increasingly replaced by or complemented with positioning technologies. Rapid advances in these technologies — with GPS (global positioning system) as the prime example — and their increasing adoption in mobile phones, have revolutionized the field causing a new flood of movement data and application possibilities (Miller, 2010). All positioning technologies work along the same principle, where the location of a moving object is registered at sequential and discrete time steps. The resulting temporal granularity can vary widely, though, from nearly continuous measurements where the location of an object is known at any given time to scenarios where there are long time gaps in between consecutive location measurements. Besides technical deficiencies, this typically occurs when a moving object is only sensed at locations where a proximity sensor is present (Versichele et al., 2012a), when a certain activity such as making a phone call (Ahas et al., 2008; González et al., 2008) or using public transport (Pelletier et al., 2011) takes place, or when energy consumption limits the temporal granularity of location measurements (Rutz and Hays, 2009). Many existing methods for movement analysis are not designed for dealing well with these temporally sparse data types, hereafter referred to as 'episodic' movement data (Andrienko et al., 2012).

Although the location of an object cannot be precisely known during a time gap without measurements, it will frequently be modeled by means of interpolation. The simplest approach is a linear (Erwig et al., 1999) interpolation, but polynomials or splines can also be used (Bartels et al., 1987). The range of movement between two known position measurements can also be more realistically modeled as an area instead of a simple line. Stochastic models such as Brownian bridge movement models (Horne et al., 2007) have received particular attention in animal movement research, but their general use remains limited to date. Alternatively, Hägerstrand's time-geographic framework (1970) represents a versatile and well-studied approach for analytically modeling spatiotemporal movement capabilities under certain constraints (Hornsby and Egenhofer, 2002; Kuijpers et al., 2010; Pfoser and Jensen, 1999). Especially the space-time prism construct, modeling all locations that can be reached between two known locations and times, has proven valuable (Miller, 2005a). Given that episodic movement data can be interpreted as sequences of known locations and times, time-geographic concepts seem directly applicable to the problem of modeling an object's potential location during a period of time where its location is unknown.

In this chapter, we focus on the network-constrained movement of objects whose location is measured only when passing a proximity-based sensor. Because of financial constraints, the number of deployed sensors will often be relatively small compared to the study area and the *actual* co-presence of moving objects at these locations will be limited in time. In order to cope with this sparse sampling of locations, a model is developed to calculate *feasible* co-presence opportunities of the moving objects in between the sensor locations. Feasible co-presence opportunities are defined here as a heuristic subset of more realistic *potential* co-presence opportunities, the latter being solely governed by the spatiotemporal constraints of proximity detections according to the time-geographic framework. More specifically, the model will be used to calculate the number of objects that could have been 'feasibly' present during a certain time window at intermediary locations based on the actual presence (detection) of moving objects at sensor locations as input. In what follows, we will focus on the aspects of time-geographic research which are most relevant to the development of such a model.

As a constriction of movement possibilities leads to more realistic analytical results, significant attention has already been devoted to dropping the unrealistic assumption of isotropic speed in unconstrained space in the traditional time-geographic framework. This was either done by constraining space-time paths and space-time prisms to networks where each edge can be assigned a static/dynamic maximum speed (Kuijpers and Othman, 2009; Miller, 1999; Neutens et al., 2008b), or by working with isotropic speed in a constrained environment (Delafontaine, 2011). Alternatively, field-based representations have also been used (Miller and Bridwell, 2009). The focus on the co-presence of groups of moving objects in our model can also draw from a recent strain of time-geographic research which deals with groups of individuals, the spatiotemporal interaction possibilities between them, and group decision-making processes. This has led to theoretical contributions on the subject (Espeter and Raubal, 2009; Kuijpers et al., 2011; Miller, 2005b; Neutens et al., 2008a, 2007), the development of toolkits for the calculation and visualization of the time-geographic interaction potential of small groups of people (Fang et al., 2011; Kang and Scott, 2007; Neutens et al., 2010) or shared-ride trip planning (Raubal et al., 2007), and the application of space-time prism intersection algorithms using empirical commuting flow data (Farber et al., 2013; Neutens et al., 2013).

We want to make a clear distinction between *potential* and *feasible* co-presence in our model, where feasible co-presence opportunities represent a subset of more likely/realistic co-presence opportunities based on the feasibility that a location can be reached by each moving object separately. This feasibility will be based on a heuristic measure, as will be discussed in section section 4.2.2 on page 72. This aspect of the model can, again, be linked to a recent trend in time-geographic research where there has been a growing effort in trying to incorporate the probability of reachability within the interior of a space-time prism. Approaches translating reachability from a binary nature (defined by the border of the space-time prism) to a probabilistic one have only started to surface recently due to its complexity

(Huisman and Forer, 2005). The distribution of visit probability within the prism is either derived by applying random walk models (Winter and Yin, 2010, 2011), or by adapting traditional kernel density estimation techniques to a time-geographic density estimation method (TGDE) where points are no longer assumed independent from each other — either in unconstrained space (Downs et al., 2011) or in a network-constrained environment (Downs and Horner, 2012; Horner et al., 2012).

Although the developed model will be able to handle any type of episodic movement data, it will be applied to a Bluetooth tracking dataset as this methodology seems to gather momentum in the relevant literature. The movements of Bluetooth-enabled mobile devices have already been registered and subsequently analyzed for various purposes ranging from mass event crowd management (Stange et al., 2011; Versichele et al., 2012a), travel time measurement of motorized traffic (Haghani et al., 2009), mobile mapping of sporting event crowds (Versichele et al., 2012b) and social studies (Eagle and Pentland, 2005).

In the next section, the model for deriving feasible co-presence opportunities will be discussed in more detail. In short, the method consists of (1) constructing discrete node-based space-time prisms for all time gaps during which the location of a moving object is unknown; (2) reducing the extent of these time prisms by pruning locations which are deemed unfeasible to reach given the spatiotemporal constraints; and (3) calculating the intersections of all feasible time prisms during a certain time window. The output of the model consists of a series of maps showing the evolution of the number of moving objects that could have been feasibly co-present over each edge of the network. The model is then applied to a large Bluetooth tracking dataset containing thousands of individual trajectories, and will generate insights into the spatiotemporal evolution of the distribution of the crowd over the entire network despite a limited number of sampling locations.

The remainder of this chapter first formalizes the research problem, and presents the proposed model and its implementation (section 4.2). Section 4.3 on page 75 describes the results from the case study. Subsequently, we further interpret the results and present the performance of the implementation in section 4.4 on page 79. In section 4.5 on page 83, finally, we give some conclusions and suggestions for further research.

4.2 Model formalization and implementation

In this section, we will first formalize the problem and subsequently describe its implementation.

4.2.1 Potential co-presence opportunities

Let a network be represented by a graph $G = (N, E)$ consisting of a set of nodes $N = \{n\}$ and a set of interconnecting edges $E = \{e\}$, $O = \{o\}$ be a population of moving objects, $S = \{s\}$ be a set of sensors with the ability to locate the proximity of moving objects on the network, and the tuple $l_s = (e, \epsilon)$ refer to the location of a sensor on an edge e of the network where $\epsilon = [0, 1]$ represents the relative location on the edge going from the starting node (0) to the ending node (1). A trajectory of a moving object o can now be described as an ordered sequence of detections: $T_o = \{d_1, d_2, \cdots, d_n\}$ where each detection is characterized by a sensor location on the network and a starting and ending time: $d_i = (o, l_i, t_{start,i}, t_{end,i})$ with $l_i \in \{l_s\}$. Alternatively, a trajectory can also be characterized by an ordered sequence of moves in between subsequent detections: $T_o = \{m_1, m_2, \cdots, m_{n-1}\}$ where a move in between two detections d_i and d_{i+1} is denoted as: $m_i = (o, l_i, l_{i+1}, t_{end,i}, t_{start,i+1})$.

Formulating it with regard to a discrete network-constrained environment, a space-time prism $\Pi_m = \{\pi_{n,o}\}$ associated with a certain move m consists of a set of time intervals or 'potential presence intervals' $\pi_{n,o}$, each describing the potential duration of presence of a moving object o at a node n of the network during that move. The boundaries of this time interval are governed by both the temporal constraints of the detections delineating the move, as well as the maximum travel speed v_e over each edge e: $\pi_{n,o} = \left[t_{end,i} + \tau_{l_i,n}, t_{start,i+1} - \tau_{n,l_{i+1}}\right]$ with $\tau_{l_i,n}$ refering to the shortest travel time from the location of the i-th detection to node n and $\tau_{n,l_{i+1}}$ refering to the shortest travel time from node n to the location of the next detection $i+1$. At each node, the time budget in between two detections b is thus distributed over travel time and presence time: $b = t_{start,i+1} - t_{end,i} = \tau_{n,o} + \pi_{n,o}$ with $\tau_{n,o} = \tau_{l_i,n} + \tau_{n,l_{i+1}}$[1]. Figure 4.1 on the following page shows a conceptual three-dimensional representation of two discrete node-based space-time prisms. The potential co-presence $\Psi_{n,W}$ at a node n with respect to a time window $W = [t_{start,W}, t_{end,W}]$ can now be calculated as the subset of all moving objects that can reach the node during the time window: $\Psi_{n,W} = \{o \in O : \pi_{n,o} \cap W \neq \emptyset\}$. Note that an object can be present in the subsets of multiple nodes during a given time window.

Depending on the context and the research question, $\#\Psi_{n,W}$ (i.e. its cardinality) can be interpreted as a more or less instantaneous co-presence of moving objects when the duration of the time window W is small relative to the speed of movement (e.g. for pedestrian movement this could be in the order of seconds) or as an indicator of potential crowdedness when wider time windows are chosen (e.g. how many objects could have been present at a location during one hour). While the implementation is inherently node-based, $\#\Psi_{n,W}$ can easily be visualized along all edges of the underlying network. The edges can be segmented into subsegments of length σ (a value that can be specified by the user) and each segment can be attributed a value through linear interpolation between the values of $\#\Psi_{n,W}$ of both end vertices of the edge[2]. A conceptual representation of the potential co-presence is also depicted in figure 4.1 on the next page.

[1]Note that we do not take any activity participation time into account as we do not presuppose any obligatory activities taking place in between anchor points.

[2]We note that linear interpolation could potentially result in unrealistic values along long network segments on the edge of the reachable subset of network segments (where one endpoint would be reachable by a number of moving objects, and the other would not be reachable by any object). The network used for this study lacks such segments, however.

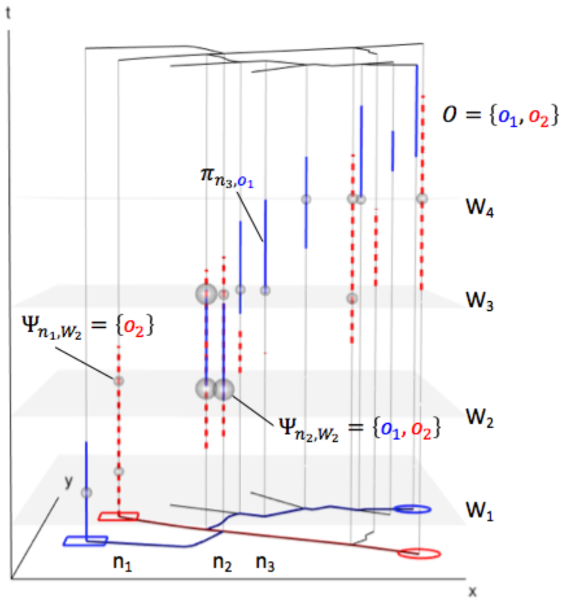


Figure 4.1: Three-dimensional representation of two node-based space-time prisms of two moving objects (1: blue, 2: red), and the resulting potential co-presences between them. Origins (squares), destinations (circles) and shortest paths for each move are plotted on the lower network projection as a reference. Vertical colored lines represent the potential presence intervals $\pi_{n,o}$ (1: solid, 2: dashed). Grey spheres at each (instantaneous) time window ($W_1 \to W_4$) represent the number of objects that could have been present at each reachable node during W ($\#\Psi_{n,W}$; small sphere: only one object, large sphere: both objects).

4.2.2 Feasible co-presence opportunities

To include the notion of feasibility of co-presence into our model, we again focus on one individual move. Despite the different approaches used to study the distribution of presence probabilities inside of a prism — ranging from random walk simulations (Winter and Yin, 2011) to kernel density estimation (Downs and Horner, 2012) — the main factor influencing this probability (in a negative way) is the travel time required to reach a certain point, meaning that longer travel times towards the border of the prism will lead to lower probabilities of presence. The way this travel cost is incorporated, however, varies. In brief, Downs and Horner (2012) calculate probabilities inside of a space-time prism by dividing the distance to reach an intermediary location by the maximum possible distance that can be covered with the available time budget b, and subsequently apply a decay function to this ratio[3]. By dividing both distances by a travel speed and using an inverse decay function, this devia-

[3]We note that this ratio is additionally modified by the dimension of the potential path tree associated with the move in order to avoid bias when combining several moves. Since our focus lies on one individual move, we can disregard this factor.

tion from the longest possible path can be formulated in the following manner: $\delta_{LP} = {}^{b}/_{\tau}$ where τ is a short notation of $\tau_{n,o}$. As $\tau \leq b$, the domain of δ_{LP} is essentially $[1, +\infty[$ with $\delta_{LP} = 1$ corresponding to a location on the border of the prism that can only be reached by spending the entire time budget travelling (without having any spare time in between) and $\delta_{LP} = max\{\delta_{LP}\}$ situated on the shortest path between both anchor points.

A possible alternative is to calculate the deviation from the shortest instead of the longest possible path. Formally, we can say that the deviation from the shortest path $\delta_{SP} = {}^{\tau_{SP}}/_{\tau}$ with τ_{SP} representing the time it takes to travel between both anchor points along the shortest path. As $\tau \geq \tau_{SP}$, the domain of δ_{SP} is $]0, 1]$ with $\delta_{SP} = 1$ corresponding to a location situated on the shortest path and $\delta_{SP} \simeq 0$ corresponding to a location on the fringe of the prism. Figure 4.2 on the following page now compares both measures for one move with a shorter and longer time budget. Both δ_{LP} and δ_{SP} have higher values along the shortest path and lower values near the border of the prism. Values for δ_{SP} are not influenced by the available time budget, although the extent of the prism clearly grows with a longer budget. In contrast, values for δ_{LP} are time-budget dependent: a larger budget will lead to larger values in the central area whereas the extreme border of the prism will be associated with values close to 1.

In order to go from potential to feasible presence opportunities, all nodes with δ_{SP} or δ_{LP} values below a certain threshold value can be pruned. The result of this heuristic operation is a space-time prism which is more constrained towards the shortest path between both anchor points. The choice between δ_{SP} and δ_{LP} as heuristic measure ultimately boils down to the question whether the probability of reachability within a space-time prism should only depend on the deviation from the shortest path (δ_{SP}) or whether the available time budget should also have an influence (δ_{LP}). In this case, δ_{LP} is proportionally related to b, so a move with a longer time budget would surpass the threshold value faster and lead to larger prisms. While the consideration of a heuristic measure will be scenario-specific, we think that δ_{SP} is more suited for the scenario that is assumed in this chapter and was already concisely described in section 4.2.1 on page 70: moving objects alternating periods of detections by sensors at specific points of interest and moves in between these locations. Provided that all points of interest attracting moving objects are covered by a sensor, we can assume that the primary purpose of the time spent in between two detections is travel. Hence we do not intend to model an object's possibility of undertaking a certain activity between two anchor points (as in most space-time accessibility frameworks). In contrast to this last approach, the probability of a moving object having passed a certain node of the network when moving in between points of interest should not be higher when the object has a larger time budget. An additional advantage of δ_{SP} is its fixed domain between 0 and 1, and its more straightforward interpretation towards calibration. More formally, we can now define the feasible co-presence on a network node as the following subset of potential co-presence: $\Omega_{n,W} = \{o \in O : \pi_{n,o} \cap W \neq \emptyset \wedge \delta_{SP}(\pi_{n,o}) \geq \delta_{SP,min}\} \subseteq \Psi_{n,W}$, with $\delta_{SP,min}$ being the threshold value imposed on δ_{SP}. The conceptual difference between

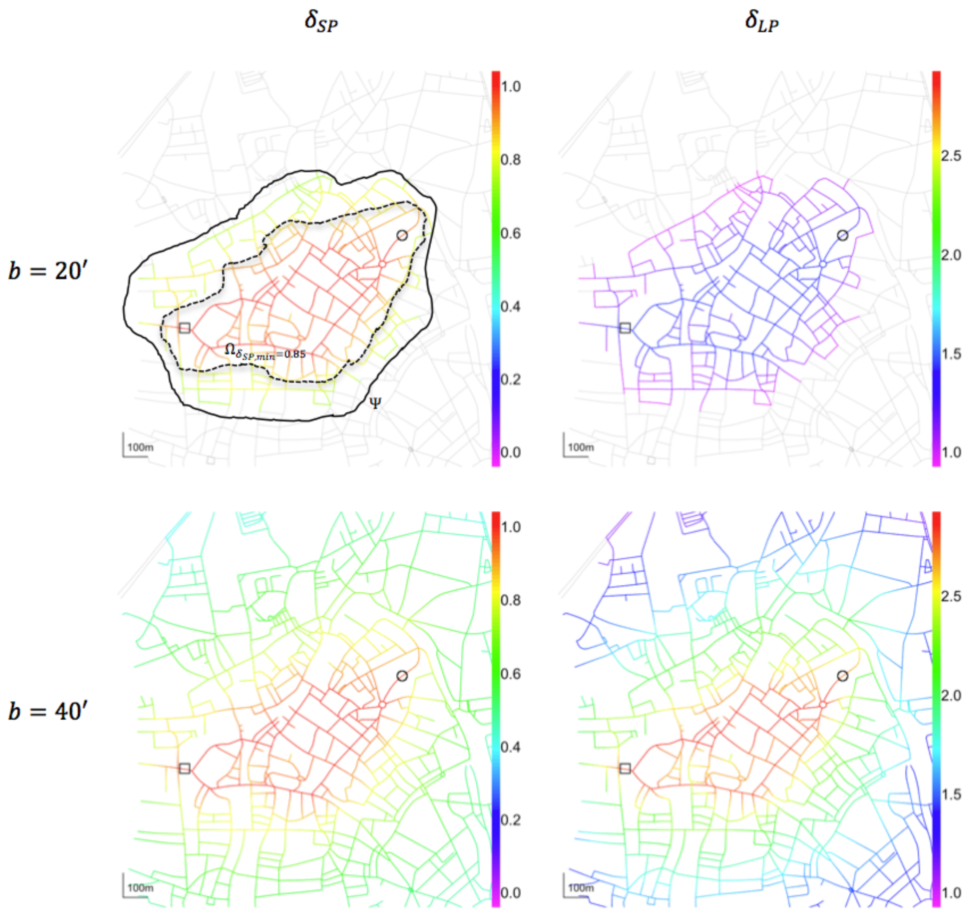


Figure 4.2: Effect of the available time budget b on the δ_{SP} and δ_{LP} measures for deriving the feasibility of presence within a space-time prism. All frames correspond to a move between the same origin (square in the south-west) and destination (circle in the north-east) with a corresponding shortest path of 1,174 meters which takes 14 minutes and 3 seconds to cross when a uniform walking speed of 5 km/h is taken into account. Grey network segments are not part of the space-time prism, the segments which are reachable are colored according to the value of δ_{SP} or δ_{LP}. The first frame show a conceptual depiction of all potentially reachable locations and the subset of feasibly reachable locations ($\delta_{SP,min} = 0.85$).

the potential reach during one move and the feasible reach is illustrated in the first frame of figure 4.2. Figure 4.3 on the next page shows the effect of the threshold value imposed on δ_{SP} in a scenario with two moving objects. A threshold of 0.95 clearly shows the fastest paths for both moves, and results in a significantly smaller interaction zone.

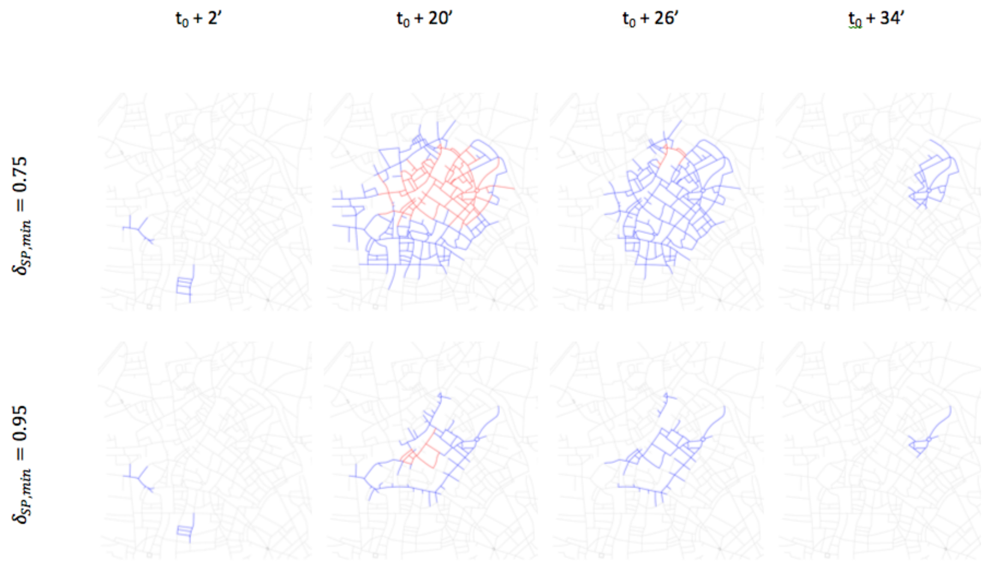


Figure 4.3: Feasible co-presence opportunities $\#\Omega_{n,W}$ for a population of 2 moving objects (1: moving from west to east, 2: moving from south to north). Both objects start their move at t_0, but object 1 has a time budget of 40 minutes whereas object 2 only has 30 minutes. A uniform walking speed of 5 km/h is assumed over the network. Blue network segments are only feasible to reach for one of the objects ($\#\Omega_{n,W} = 1$), red segments for both objects ($\#\Omega_{n,W} = 2$).

4.2.3 Implementation

The model described in this section was implemented as an extension to a previously implemented spatiotemporal accessibility toolkit (Neutens et al., 2010). Briefly summarized, this *Java* based standalone toolkit is able to assess the spatiotemporal accessibility of urban opportunities to one or multiple persons based on information about the transportation system, urban opportunities and individual activity schedules. An extension of the original toolkit was necessary for the handling of Bluetooth tracking data, and the implementation of a new data structure for modeling feasible co-presence opportunities. The visualization of this data structure was done in the *R* suite (2.14.0), where a function was implemented that can automatically generate time-series of maps based on the output of the toolkit. An evaluation of the performance of the toolkit will be given in section section 4.4.2 on page 82.

4.3 Case study: ‘Ghent Light Festival 2012’

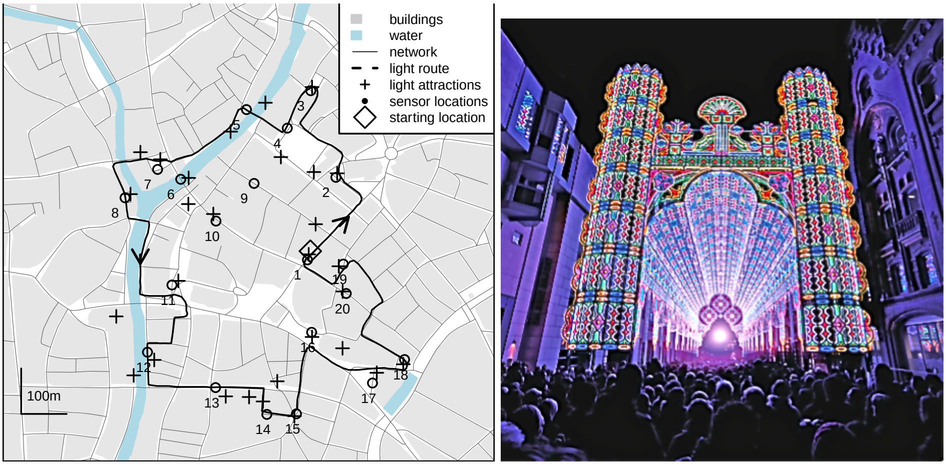
In order to test the model developed in section 4.2 on page 70 and its implementation, we will apply it to a large (episodic) Bluetooth tracking dataset and focus on the network-constrained movements of a crowd in between sensor locations based on location measurements at the sensor locations. The Bluetooth tracking dataset in this chapter was gathered during the 'Ghent Light Festival 2012'[4]. During this four day long event (26–29 January), 29 artworks making use of light were dispersed over the city center of Ghent, Belgium. Every night, these artworks were functioning from 6 pm until midnight, and visitors could walk along them following a 'light route' of 4 kilometers. Over the four days, an estimated 500,000 visitors attended the festival for free. In order to get a clearer understanding of the spatiotemporal dynamics of the crowd, 25 Bluetooth scanners were strategically placed along the light route in close vicinity of the artworks where possible (sometimes a very close distance was not possible due to the unavailability of a secure location for the scanner or a power source). For more technical details on the Bluetooth scanners and sensors, we refer the reader to (Versichele et al., 2012a). At most locations, class 2 Bluetooth sensors were deployed. If the area to be covered was larger (e.g. at squares) or the sensor needed to be deployed on a higher floor of a building, we used more sensitive class 1 sensors with a larger reach (i.e. locations 1, 3, 4, 7, 11 and 16 on figure 4.4 on the facing page). The scanners sensed the presence of Bluetooth-enabled mobile devices whenever they were nearby. Over the course of four days, 35,207 unique Bluetooth devices were registered during the running hours of the event. These devices were responsible for 38,503 trajectories (after splitting of the original trajectories at time gaps of at least six hours in length in order to distinguish between several visits). Five scanners were dropped from the remainder of the analysis (1 due to technical issues, 4 because they were located indoors instead of outdoors and we intend to model feasible co-presences on an outdoor network). Figure 4.4 on the next page shows the city center of Ghent, the 'light route', the locations of the light attractions and the Bluetooth scanners. It also shows a photograph of the artwork 'Luminarie de Cagna' located at the starting location of the light route. The network used as input to the implementation consisted of 1,381 nodes and 1,880 interconnecting edges. A uniform maximum travel (walking) speed of 5 km/h was used for all edges of the network (Bohannon, 1997).

In the remainder of this section, two analyses will be performed to illustrate the potential of our approach. First, we will focus on a subpopulation of the total tracked population in order to visualize and interpret the movements of 'early-bird' visitors (i.e. visitors present at the starting location at 6 pm, O_1) during the third day of the event. Next, we will expand the analysis to include the movements of the entire population of visitors on the same day (O_2).

4.3.1 Early-birds

Only the visitors that were detected at the starting location of the route between 5:50 pm and 6:05 pm on day 3 (6 pm being the time the artworks were illuminated) were selected from the total population. This analysis will now focus on these 388 'early-bird' visitors (symbolized as O_1) and how they move around in the study area. The feasible co-presence opportunities were calculated as explained in section 4.2 on page 70. The pruning to feasible space-time

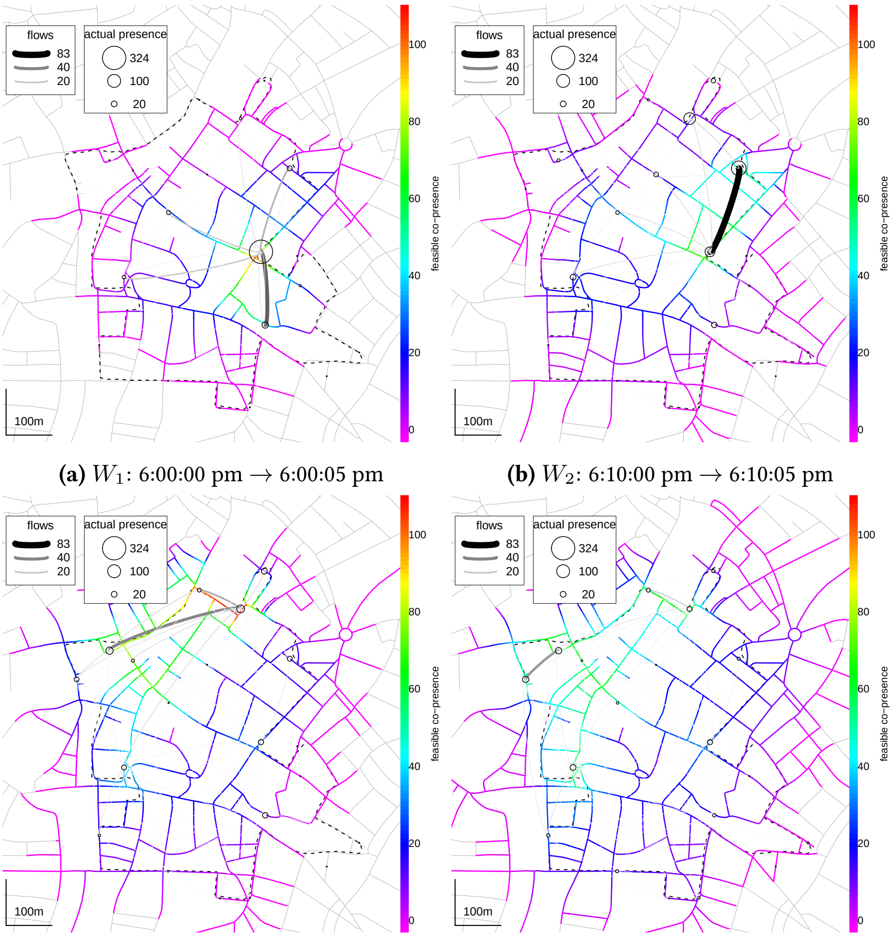
[4] www.lichtfestivalgent.be



(a) **(b)**

Figure 4.4: (a) Study area of the 'Ghent Light Festival 2012'. Arrows indicate the direction of the programmed light route, numbers refer to the sensor locations. (b) View at the starting location, looking in a north-eastern direction.

prisms was done with a threshold value of $\delta_{SP,min} = 0.75$. Time windows of five seconds in length were calculated every five minutes. Six characteristic snapshots are shown in figure 4.5 on the next page. Besides feasible co-presence (represented by a color scale over the network according to $\#\Omega_{n,W}$, with grey segments not reachable by any visitor), the moves associated with the calculated space-time prisms are also visualized in an aggregated way. The width of the lines representing these flows varies according to the number of moves in the flow that overlap with the time window of the snapshot. The lines also curve to the right going from origin to destination in order to make flow directions easier to interpret. Finally, the number of visitors that was actually detected (over five minutes) is also depicted at each sensor location by proportionally sized circles. At 6 pm (W_1), visitors have clearly assembled at the starting location of the light route. This is both supported by the actual co-presence of devices detected between 6 pm and 6:05 pm (324, highest in the entire time series), and the feasible co-presence of almost 100 devices on the closest node. As can be seen by all the aggregated moves pointing towards it, visitors seem to be still gathering at the starting location. Ten minutes later (W_2), the opposite is visible: all moves point away from the starting location, with one flow in particular dominating towards the second attraction in the north-eastern part of the light route. Correspondingly, the feasible co-presence is also highest in this region although somewhat lower than in the previous time window. Twenty minutes later (W_3), we can see that the bulk of the visitors is now concentrated in the northern part of the route. The number of feasibly co-present devices is now at its maximal value over the entire time series (106). Again twenty minutes later (W_4), we see a diffusion of the group further along the route while the tail of the crowd seems to be trailing in the north. The next window (W_5) is similar in appearance with a slightly more pronounced concentration now in the west. At 8 pm (W_6), finally, what is left of the original group is spread out over most of the center and no real concentration is visible anymore.



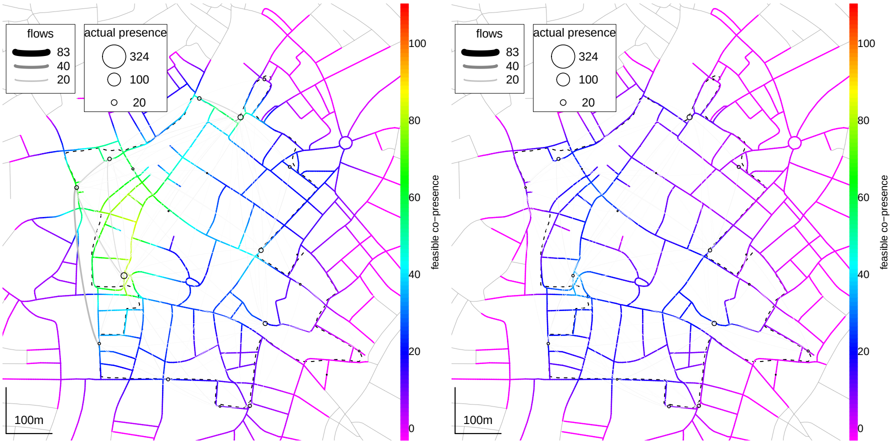
(a) W_1: 6:00:00 pm $\longrightarrow$ 6:00:05 pm

(b) W_2: 6:10:00 pm $\longrightarrow$ 6:10:05 pm

(c) W_3: 6:30:00 pm $\longrightarrow$ 6:30:05 pm

(d) W_4: 6:50:00 pm $\longrightarrow$ 6:50:05 pm

Figure 4.5: Visualization of 'early-bird' ($\#O_1 = 388$) *feasible* co-presences $\#\Omega_{n,W}$ ($\delta_{SP,min} = 0.75$) by a color gradient over the road network. Six characteristic snapshots ($|W| = 5s$) during the course of the third evening of the 'Ghent Light Festival' are shown ($W_1 \rightarrow W_6$). Curved arcs depict flows in between scanner locations during the time window (curving to the right from origin to destination). Circles represent the actual presence of detected devices over 5 minutes ($[t_{start,W}, t_{start,W} + 5min]$) at each scanner location.



(e) W_5: 7:00:00 pm → 7:00:05 pm **(f)** W_6: 8:00:00 pm → 8:00:05 pm

Figure 4.5: Continued.

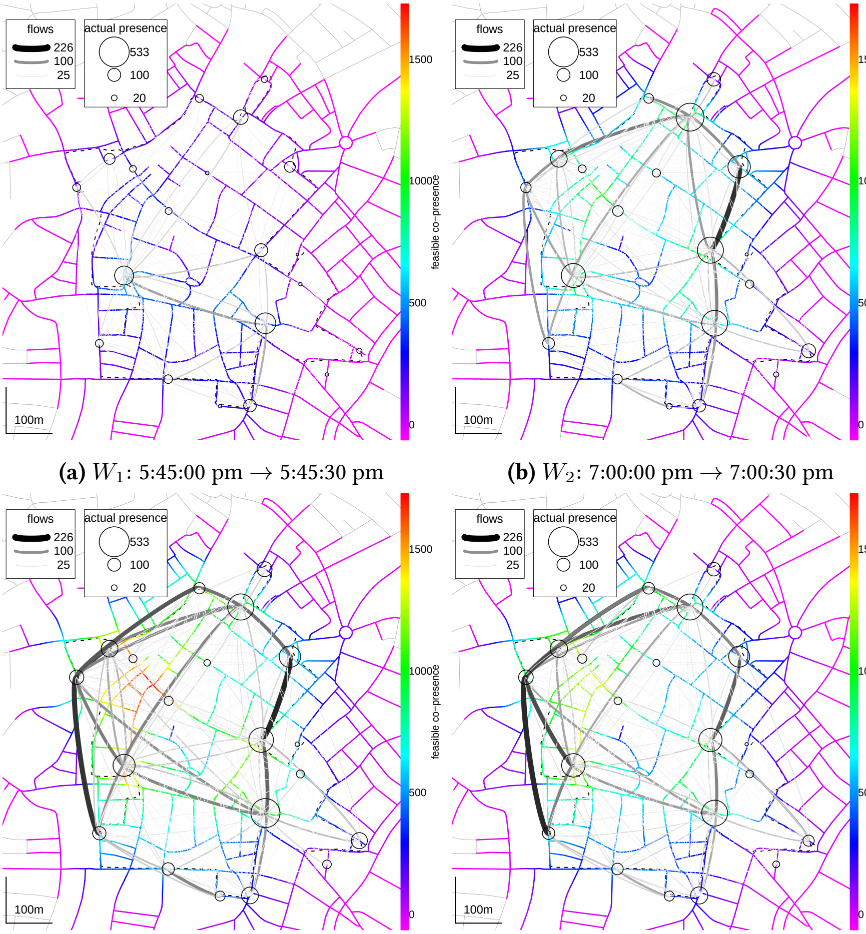
4.3.2 All visitors

Subsequently, the trajectories generated by all visitors during day 3 of the festival ($\#O_2 = 15,564$) were used as input to the model, and a similar analysis as in section 4.3.1 on page 76 was performed. Four snapshots are shown in figure 4.6 on the next page. Fifteen minutes before the start of the event (W_1), little movement is recorded and feasible co-presences are correspondingly low as well. One hour after the start (W_2), the aggregated moves indicate that the crowd is moving along the route. The feasible co-presence of visitors is more or less evenly spread over large parts of the route. Two hours later (W_3), visitor flows have grown in magnitude and a zone with significantly higher feasible co-presences has become visible around the intersection of the south-oriented flows in the west of the study area. At 10 pm (W_4), visitor activity has clearly decreased. The highest feasible co-presences are still located in the west, but are already significantly lower than one hour earlier.

4.4 Discussion

4.4.1 Interpretation of feasible co-presence and its relation to actual co-presence

As demonstrated in section 4.3.1 on page 76 for a subpopulation and in section 4.3.2 for the entire population of visitors to the 'Ghent Light Festival 2012', our model is able to derive feasible co-presence opportunities from large amounts of episodic movement data. The output of the model was used for visualizing the estimated distribution of a crowd over an entire network based on a relatively small (i.e. 25) number of sampling points. As a reference, the



(a) W_1: 5:45:00 pm $\rightarrow$ 5:45:30 pm **(b)** W_2: 7:00:00 pm $\rightarrow$ 7:00:30 pm

(c) W_3: 9:00:00 pm $\rightarrow$ 9:00:30 pm **(d)** W_4: 10:00:00 pm $\rightarrow$ 10:00:30 pm

Figure 4.6: Visualization of *feasible* co-presences $\Omega_{n,W}$ ($\delta_{SP,min} = 0.75$) for all detected visitors ($\#O_2 = 15,564$) by a color gradient over the road network. Four characteristic snapshots ($|W| = 30s$) during day 3 of the 'Ghent Light Festival' are shown ($W_1 \rightarrow W_4$). Curved arcs depict flows in between scanner locations during the time window (curving to the right from origin to destination). Circles represent the actual presence of detected devices over 5 minutes ($[t_{start,W}, t_{start,W} + 5min]$) at each scanner location.

numbers of devices detected over five minutes were additionally depicted for each sensor location in figure 4.5 on page 78 and figure 4.6 by means of proportionally sized circles. It is important, however, to make a clear distinction between both variables and to understand why they are not directly related to each other. The number of devices detected at a certain location is an expression of *actual* co-presence (i.e. the number of visitors actually present at a sensor location during a certain time period), whereas *feasible* co-presence represents an estimation of how many visitors could have feasibly passed an intermediary location on the network during a certain time period. As such, actual co-presence serves as an input of the model and feasible co-presence serves as its output. Additionally, it is important to note that an object can only contribute to feasible co-presence if it is moving between two sensor locations. This distinction is especially important when trying to translate model results to crowdedness levels. Feasible co-presence levels over the edges of the network do not represent actual crowdedness as such, but are an estimation of the number of visitors 'on the move' that could have passed through each edge going from one location to another. In an extreme case, a completely static crowd would not deliver any co-presence opportunities whereas certain (sensor) locations could certainly be classified as crowded.

To gain more insight into this distinction, the relation between both variables has been plotted in figure 4.7 on the following page for two locations with sensors that were incorporated into the model and one location with a sensor that was not used in the model. Although limited to three locations, the feasible co-presence values seem larger than co-located actual presence values. This is due to feasible co-presence values close to the location of a specific sensor not only being affected by displacements starting or ending at that sensor but also between neighboring sensors. The larger overestimation at sensor location 7 is most probably due to the intersection of different moves in between the more nearby sensors in contrast with sensor location 2, which is more isolated. This effect was also visible in figure 4.6 on the preceding page. As a first and preliminary step in validating the model, the relation between both variables was also plotted for one location with a sensor that was not used in the model. The trend seems similar to the one for location 7. It must be stressed, however, that this only comprises a very concise attempt at validating the model and further validation is certainly necessary. One way could be to more systematically exclude sensor locations from the model and studying their relation between measured actual co-presence and estimated feasible co-presence as in figure 4.7 on the following page. The inclusion of more ground-truth data (e.g. GPS devices or cameras for estimating the crowdedness levels on certain network segments) will be particularly helpful in such an endeavor. A systematic calibration of the model parameters (mainly $\delta_{SP,min}$) is also necessary for enhancing the reliability and interpretability of the model outputs. The threshold of $\delta_{SP,min}$ will certainly depend on the context. In this case study, we set the threshold at 0.75 in order to incorporate the pre-planned light trajectory (which more or less connects the light works by shortest paths) but at the same time leave some room for movements deviating from this trajectory. In case this trajectory would have actually been physically imposed on the crowd (e.g. by fences), we could have set the threshold closer to 1. In contrast, larger degrees of freedom for both movement and dwelling behavior (e.g. in case there would be no pre-planned trajectory) would be modeled by lower $\delta_{SP,min}$ values.

The selective modeling of (moving) feasible co-presence instead of (static) actual presence might open up new and interesting avenues of research as there usually is a positive relation between crowd movement and friction and the probability of stampeding and tram-

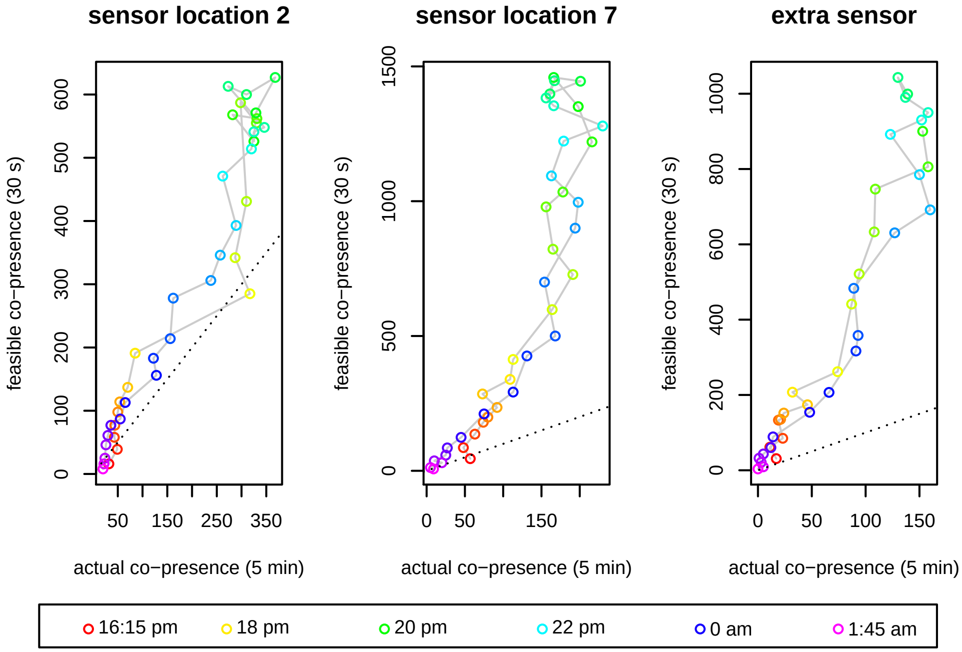


Figure 4.7: Relation between feasible and actual co-presence of all visitors (O_2) at two sensor locations used in the model and one extra sensor that was not used in the model (situated in between sensors 5 and 7 on the light route). The numbers of the sensor locations refer to those in figure 4.4 on page 77. The color of the points varies according to a time scale. Dotted lines represent $y = x$.

pling (Helbing et al., 2000). Our approach could be used to either detect (as in figure 4.6 on page 80) or predict (e.g. by simulations based on previously gathered tracking data) locations where large groups could have been or could be feasibly co-present, thus causing elevated crowd risks. As with any other tracking technology, specific attention should be given to the representativeness of the gathered movement data. In this case study, the 38,503 registered trajectories represent almost 8% of the total estimated visitor population. As such, one should be careful when extrapolating findings to the entire visitor population. On the other hand, the gathered dataset covers a significantly larger share of individuals in comparison with other traditional techniques such as the use of GPS loggers and/or questionnaires, and in a higher spatial accuracy than cell phone tracking methods.

4.4.2 Performance of the toolkit

A brief summary of the performance of the toolkit and its scalability with regards to the number of moving objects is shown in figure 4.8 on the next page. Calculations were done on a non-dedicated system with a 2.66 Ghz Intel Core 2 Duo CPU and 8 Gb of internal memory. The software environment consisted of Mac OS X 10.7.5, *R* 2.14.0 and *Java* SE 6. The scenario used for the calculation was the same as in section section 4.3.2 on page 79, and the sample size was varied by subsampling the original population. While calculation times for large populations are still acceptable for post-event analyses, real-time processes will clearly necessitate improvements in the efficiency of the implementation. One possible avenue is the parallelization of the calculation process. Dynamic instead of static maximum speeds over the network segments could be another interesting addition to the toolkit.

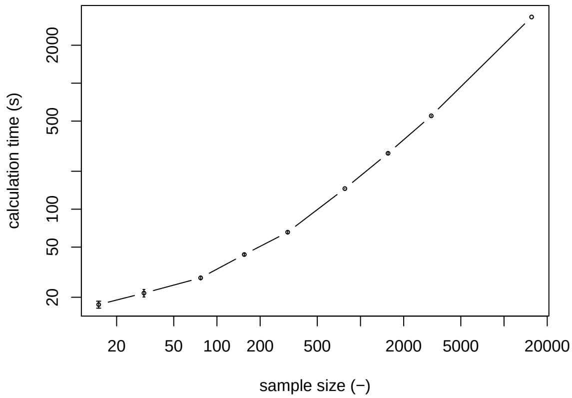


Figure 4.8: Relation between sample size and calculation time as a measure of the performance and scalability of the implementation. Means as well as standard deviations are shown ($n = 3$).

4.5 Conclusions

This study started with a discussion on the nature of 'episodic' movement data, and how traditional methods are poorly equipped for the reconstruction of an object's movement during time gaps where its position is unknown. A concise discussion on time geography and its potential value in addressing this issue followed, mainly focusing on three important research trends which can be identified in the available literature: the translation of conceptual time-geographic concepts to realistic (e.g. network-constrained) environments, the valuation of reachability within a space-time prism, and the focus on multiple individuals and their interaction spaces instead of individuals.

Simultaneously picking up on these trends, a model was developed for the derivation of feasible co-presence opportunities for groups of moving objects whose location is only registered at certain sensor locations on a road network. This model was subsequently implemented as an extension on a previously implemented toolkit, and illustrated on a Bluetooth tracking dataset gathered during the 'Ghent Light Festival 2012' ($\pm$ 500,000 visitors over four days). Feasible co-presences at regular time windows were calculated at each network node and subsequently interpolated over all network segments for a subpopulation and the entire population of tracked visitors on the third event day. This way, the estimated distribution of both groups over time was visualized and subsequently interpreted in the context of the mass event. Special attention was also given to both the relation between and the important distinctions between feasible and actual presence when interpreting output results.

Although the model can be considered a promising approach to dealing with the absence of positional knowledge in episodic data — in the sense that the generated maps reveal patterns that correlate with the expected outcome and that they are interpretable given the context of the event — more focus should be devoted to a further validation and calibration of the model with additional ground-truth data. As this data was lacking in the study, this was not possible at present but should certainly receive further attention in the near future. The stability of the model and its output with regards to the inclusion or exclusion of different sensors from the dataset should also be investigated further. Additional contextual knowledge can also be incorporated into the model by varying walking speeds over different network segments (either in a static way or dynamically depending on the crowdedness) instead of assuming a fixed walking speed of 5 km/h. Finally, the toolkit was evaluated to be sufficiently efficient for current offline analyses but improvements will be necessary for real-time calculations.

References

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., and Hecker, D. (2012). Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. *KI - Künstliche Intelligenz*, 26(3):241–251.

Axhausen, K., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.

Bartels, R., Beatty, J., and Barsky, B. (1987). *An introduction to splines for use in computer graphics and geometric modeling.* Morgan Kaufmann Publishers, Los Altos, CA.

Bohannon, R. W. (1997). Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age and ageing*, 26(1):15–19.

Delafontaine, M. (2011). Modelling potential movement in constrained travel environments using rough space–time prisms. *International Journal of Geographical Information Science*, 25(9):1389–1411.

Downs, J. and Horner, M. (2012). Probabilistic potential path trees for visualizing and analyzing vehicle tracking data. *Journal of Transport Geography*, 23:72–80.

Downs, J., Horner, M., and Tucker, A. (2011). Time-geographic density estimation for home range analysis. *Annals of GIS*, 17(3):163–171.

Eagle, N. and Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Erwig, M., Ting, R. G., Schneider, M., and Vazirgiannis, M. (1999). Spatio-Temporal Data Types : An Approach to Modeling and Querying Moving Objects in Databases. *GeoInformatica*, 3(3):269–296.

Espeter, M. and Raubal, M. (2009). Location-based decision support for user groups. *Journal of Location Based Services*, 3(3):165–187.

Fang, Z., Tu, W., Li, Q., and Li, Q. (2011). A multi-objective approach to scheduling joint participation with variable space and time preferences and opportunities. *Journal of Transport Geography*, 19(4):623–634.

Farber, S., Neutens, T., Miller, H. J., and Li, X. (2013). The Social Interaction Potential of Metropolitan Regions: A Time-Geographic Measurement Approach Using Joint Accessibility. *Annals of the Association of American Geographers*, 103(3):483–504.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hägerstrand, T. (1970). What about people in Regional Science? *Papers in Regional Science*, 24(1):6–21.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., and Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160:60–68.

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490.

Horne, J. S., Garton, E. O., Krone, S. M., and Lewis, J. S. (2007). Analyzing animal movements using Brownian bridges. *Ecology*, 88(9):2354–2363.

Horner, M. W., Zook, B., and Downs, J. a. (2012). Where were you? Development of a time-geographic approach for activity destination re-construction. *Computers, Environment and Urban Systems*, 36(6):488–499.

Hornsby, K. and Egenhofer, M. J. (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):177–194.

Huisman, O. and Forer, P. (2005). The complexities of everyday life : balancing practical and realistic approaches to modelling probable presence in space-time. In *Proceedings of the 17th Annual Colloquium of the Spatial Information Research Centre (SIRC)*, pages 155–168.

Kang, H. and Scott, D. M. (2007). An integrated spatio-temporal GIS toolkit for exploring intra-household interactions. *Transportation*, 35(2):253–268.

Kuijpers, B., Grimson, R., and Othman, W. (2011). An analytic solution to the alibi query in the space–time prisms model for moving object data. *International Journal of Geographical Information Science*, 25(2):293–322.

Kuijpers, B., Miller, H., Neutens, T., and Othman, W. (2010). Anchor uncertainty and space-time prisms on road networks. *International Journal of Geographical Information Science*, 24(8):1223–1248.

Kuijpers, B. and Othman, W. (2009). Modeling uncertainty of moving objects on road networks via space-time prisms. *International Journal of Geographical Information Science*, 23(9):1095–1117.

Miller, H. (1999). Measuring Space-Time Accessibility Benefits within Transportation Networks: Basic Theory and Computational Procedures. *Geographical Analysis*, 31(1):1–26.

Miller, H. (2005a). A Measurement Theory for Time Geography. *Geographical Analysis*, 37(1):17–45.

Miller, H. (2005b). Necessary space - time conditions for human interaction. *Environment and Planning B: Planning and Design*, 32(3):381–401.

Miller, H. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201.

Miller, H. and Bridwell, S. (2009). A Field-Based Theory for Time Geography. *Annals of the Association of American Geographers*, 99(1):49–75.

Millonig, A. and Gartner, G. (2008). Shadowing – Tracking – Interviewing: How to Explore Human Spatio-Temporal Behaviour Patterns. In Gottfried, B. and Aghajan, H., editors, *Proceedings of the 2nd Workshop on Behaviour Monitoring and Interpretation (BMI '08)*, volume 396, pages 1–14, Kaiserslautern.

Neutens, T., Farber, S., Delafontaine, M., and Boussauw, K. (2013). Spatial variation in the potential for social interaction: A case study in Flanders (Belgium). *Computers, Environment and Urban Systems*, 41:318–331.

Neutens, T., Schwanen, T., Witlox, F., and De Maeyer, P. (2008a). My space or your space? Towards a measure of joint accessibility. *Computers, Environment and Urban Systems*, 32(5):331–342.

Neutens, T., Van de Weghe, N., Witlox, F., and De Maeyer, P. (2008b). A three-dimensional network-based space–time prism. *Journal of Geographical Systems*, 10(1):89–107.

Neutens, T., Versichele, M., and Schwanen, T. (2010). Arranging place and time: A GIS toolkit to assess person-based accessibility of urban opportunities. *Applied Geography*, 30(4):561–575.

Neutens, T., Witlox, F., and Demaeyer, P. (2007). Individual accessibility and travel literature review on time geography possibilities : A literature review on time geography. *European Journal of Transport Infrastructure and Research*, 7(4):335–352.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Pfoser, D. and Jensen, C. (1999). Capturing the Uncertainty of Moving-Object Representations. In Güting, R. H., Papadias, D., and Lochovsky, F., editors, *Advances in Spatial Databases*, volume 1651 of *Lecture Notes in Computer Science*, pages 111–131. Springer.

Raubal, M., Winter, S., Teβmann, S., and Gaisbauer, C. (2007). Time geography for ad-hoc shared-ride trip planning in mobile geosensor networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(5):366–381.

Rutz, C. and Hays, G. C. (2009). New frontiers in biologging science. *Biology letters*, 5(3):289–92.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012a). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2):208–220.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., and Van de Weghe, N. (2012b). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10):14196–14213.

Winter, S. and Yin, Z.-C. (2010). Directed movements in probabilistic time geography. *International Journal of Geographical Information Science*, 24(9):1349–1365.

Winter, S. and Yin, Z.-C. (2011). The elements of probabilistic time geography. *GeoInformatica*, 15(3):417–434.

5

Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: a case study of Ghent, Belgium

Versichele, M., De Grootte, L., Claeys Bouuaert, M., Neutens, T., Moerman, I., Van de Weghe, N.

under review

Abstract *The rapid evolution of information and positioning technologies, and their increasing adoption in tourism management practices allows for new and challenging research avenues. This chapter presents an empirical case study on the mining of association rules in tourist attraction visits, registered for fifteen days by the Bluetooth tracking methodology. This way, this chapter aims to be a methodological contribution to the field of spatiotemporal tourism behavior research by demonstrating the potential of ad-hoc sensing networks in the non-participatory measurement of small-scale movements. An extensive filtering procedure is followed by an exploratory analysis, analyzing the discovered associations for different visitor segments and additionally visualizing them in 'visit pattern maps'. Despite the limited duration of the tracking period, we were able to discover interesting associations and further identified a*

tendency of visitors to rarely combine visits in the center with visits outside of the city center. We conclude by discussing both the potential of the employed methodology as well as its further issues.

5.1 Introduction

Due to the complex nature of tourism, tourism management endeavours increasingly revolve around studying how tourists behave on an individual level (McKercher, 1999). The spatiotemporal behavior of tourists is of a particular importance as a tourist trip is essentially the result of the balancing of a certain time budget with personal preferences for visiting certain attractions or destinations which are geographically separated (Lew and McKercher, 2006). Despite this importance, empirical studies into tourist mobility have traditionally been rather scarce due to the labor-intensive and often expensive nature of traditional methods such as direct observation (Hartmann, 1988) or personal interviews (Kemperman et al., 2009). Space-time diaries (Connell and Page, 2008; Janelle et al., 1988; Lau and McKercher, 2006) shift some of the weight away from the researchers but are often characterized by a low reliability as respondents tend to forget or neglect to register certain activities. Recently, however, tracking technologies offer a more scalable and objective way to capture spatiotemporal behavior in a detailed way (Shoval and Isaacson, 2009). The use of global navigation satellite systems — such as GPS — is currently the dominant approach and its adoption in tourism research through the distribution of logging devices is well-documented (Shoval and Isaacson, 2007a,b; Shoval et al., 2011; Tchetchik et al., 2009). An alternative approach is to track the movement of mobile phones through a cell tower network without the direct participation of the phone's owner (González et al., 2008; Ratti et al., 2006). Particularly in Estonia, this method has already been extensively used for studying regional movement patterns of tourists (Ahas et al., 2007a, 2008).

Despite the undeniably important contribution of both tracking methodologies to the research field, we argue that both approaches have certain limits. The distribution of logging devices necessitates the direct collaboration of the tracked individual. This makes it hard to scale up the methodology to large groups of individuals. Additionally, any participatory methodology presents a risk for self-selection bias where individuals with certain characteristics would show a higher degree of cooperation and thus be overrepresented in the sample. Finally, GPS technology is not applicable to indoor contexts. Cell phone tracking, on the other hand, encompasses other limitations. First, the spatial accuracy of the method is limited by the density of cell towers over the study area. In Estonia, for example, around 50% of measurements were correct to within only 400 meters in urban areas and only 2600 meters in rural areas (Ahas et al., 2007b). While this does not hinder the study of regional movements, it does pose a problem when studying movement within a certain tourist destination (e.g. a city). Second, these datasets are property of mobile operators and — as such — not freely available. In summary, it seems that small-scale spatiotemporal behavior cannot be measured without the direct involvement of the individual to be tracked. This hinders studying larger groups of individuals.

A recent alternative in the non-participatory tracking of mobile phones is the use of ad-hoc sensor networks distributed over a study area. Bluetooth technology, for example, has already been employed for studying pedestrian flows at mass events (Delafontaine et al., 2012; Stange et al., 2011; Versichele et al., 2012a,b) and in social studies (Eagle and Pentland, 2005). WiFi (Bonné et al., 2013) and RFID (Öztayşi et al., 2009) technology provide similar possibilities. Due to the limited coverage of each sensor, a careful deployment of sensors may thus provide movement records with a granularity that is much smaller than the accuracy level of cell phone tracking data. Despite their potential in the non-participatory registration of small-scale movements, we have as yet no indication of their use for tourism management purposes.

This chapter aims to address this issue by presenting a case study where visitors to tourist attractions in Ghent, Belgium were registered through an ad-hoc Bluetooth sensor network. Due to the novelty of Bluetooth technology — and the use of ad-hoc sensing networks in general for that matter — we will not only elaborate extensively on the working principle of the methodology, but also on the analytical potential of such tracking data. Ad-hoc sensor network data lack the typical socio-demographic or psychographic variables used as explanatory factors in various studies related to tourism behavior. In contrast with hypothesis testing procedures, sensor network data often need to be investigated without any a priori assumptions. The collection of such methods that can be used to discover (non-trivial) patterns and knowledge from large data sets is called data mining (Fayyad et al., 1996).

The application of data mining techniques to tourism datasets is not new. In this section, we give a short overview of prior work that has used data mining in the context of tourism research. *Regression* techniques have frequently been used to model tourism demand (Song and Li, 2008; Witt and Witt, 1995), either through time-series approaches (Burger et al., 2001; Hong et al., 2011; Law, 2000; Palmer et al., 2006) or causal approaches linking the demand to external variables (Law and Au, 1999). *Clustering* or *segmentation* — which redistributes data points into clusters of higher similarity — has been employed in identifying different tourist types or profiles (Bloom, 2005; Cini et al., 2010; Dolničar, 2004; Dolničar and Leisch, 2003; Tchetchik et al., 2009). *Sequential pattern mining* on visitor trajectories enables finding patterns in the order in which certain activities or visits take place (Orellana et al., 2012; Shoval and Isaacson, 2007a). *Classification* — which can be interpreted as the categorical variant of regression — has been applied in a diverse number of contexts such as tourist (Law, 2000) and business traveler (Law et al., 2006) expenditure, and hotel customer profiling (Min et al., 2002). Finally, *association rule learning* is concerned with discovering associations between variables without fixing the output variable, as is the case in classification. In comparison to the other techniques, implementations of association rule learning in tourism research are rather scarce. Documented applications found in literature include tourism product development (Al-Salim, 2008; Liao et al., 2010), domestic tourist profiling (Emel et al., 2007), analysis of behavior on touristic websites (Rong et al., 2012), and change and trend identification in Hong Kong outbound tourism (Law et al., 2011).

This chapter aims to be a methodological contribution to the field of spatiotemporal tourism behavior research by demonstrating the potential of ad-hoc sensing networks in the non-participatory measurement of small-scale movements. We describe a case study where visitors to 14 tourist attractions were registered through Bluetooth technology sensing the mobile devices they were carrying around. In an attempt to investigate the analytical potential of the resulting data, we employ an association rule learning algorithm to mine for interesting patterns in the combinations of visits to different attractions. As the tracking data are completely anonymous, it is impossible to directly distinguish between local visitors and actual tourists as defined by the World Tourism Organization: people *"traveling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes"* (World Tourism Organization, 1995). By deploying sensors in 14 hotels, however, some visitors will be identified as hotel guests therefore giving a strong suggestion that they are indeed tourists. Extra context is added by tracking visitors at the tourist inquiry desk as well. Combining the tracking data with these contextual assumptions, we will investigate patterns for different visitor segments (e.g. those that were only detected on one day, those that were identified as hotel guests, etc.). For the sake of clarity, we will always use the term visitors instead of further labeling them as tourists.

The remainder of the chapter is organized as follows. In section 5.2, we first discuss the Bluetooth tracking methodology and its specific implementation in the case study (section 5.2.1). Next, we describe association rule learning in more detail (section 5.2.2 on page 95) and how the information it generates can be summarized in 'visit pattern maps' (section 5.2.3 on page 96). Section 5.3 on page 97 outlines the filtering of the raw tracking data in detail, and section 5.4 on page 99 presents a first data exploration. The actual association rule mining is performed for the different visitor segments in section 5.5 on page 105. We finish with a discussion and conclusion in section 5.6 on page 108.

5.2 Methods and data

5.2.1 Bluetooth tracking

For this study, scanners with Bluetooth sensors were deployed at 29 locations in and around the historical center and the 'arts quarter' of Ghent (Belgium) for fifteen days in May of 2012. Ghent was chosen as the study area because of its rather unique touristic character: despite its wealth of attractions and historical significance, it was once described as *'Belgium's best kept secret'* (Lonely Planet, 2011) due to the nearby presence of better known destinations such as Bruges. As a result, it attracts a more diverse (and probably less predictable) audience. Coupled with the fact that Ghent also serves as a university city, and that its historical center also serves a residential function, the city represents an intriguing yet challenging test bed for the tracking methodology. Additionally, some of its attractions are located at considerable distances from the historical center and the tourism department was very receptive to any methodology which could offer additional insights in the visiting behavior over the entire city.

An overview of the study area and the sensor locations is given in figure 5.1 on the next page. The full names of the different venues are shown in figure 5.4 on page 101. The locations consist of fourteen hotels (a–n), three open (1–3) and eleven closed (4-14) tourist attractions, and the inquiry desk for tourists. We make the distinction between open and closed attractions based on the need for visitors to either buy a ticket or register. The open (i.e. no registration required) attractions consisted of a cathedral, a church and an indoor market. All closed attractions were museums, covering a wide range of interests such as classic/modern arts, history, textiles, and the former 'Castle of the Counts'. In 2012, these fourteen selected attractions were responsible for around 76% of the total number of visits to all attractions in Ghent. The hotels comprised the entire range of common classes and price-ranges: one hotel without stars (a), one *- (b), two **- (c–d), four ***- (e–h), four ****-hotels (i–l), and two hostels (m–n). Together, these fourteen hotels contained 67% of the total number of available beds in the city.

The Bluetooth scanners continuously searched for discoverable Bluetooth devices within their detection range, and registered the MAC address and COD (class of device) code of each detected device together with the detection timestamp. The MAC address acts as a unique identifier of the detected device, and can be used to link different detections (at different locations) to the same device and thus generate trajectories. The COD code can be used to deduce the type of device (phone, car kit, mp3-player, etc.). More details on the Bluetooth tracking methodology and the deployed hardware can be found in a previous study (Versichele et al., 2012a). All Bluetooth sensors used were class 2 devices, which have a theoretical communication range of around 10 meters according to the official Bluetooth specifications. The actual detection range of Bluetooth sensors, however, largely depends on the environment and the presence/absence of a line-of-sight between the sensor and the detected device. As such, the exact location of each sensor was chosen by balancing the need for an optimal position (in order to only detect devices inside the attraction or hotel) and the sake of convenience (i.e. the presence of a power supply). All Bluetooth scanners were connected to the internet (18 wired, 11 wireless) in order to facilitate the remote monitoring of their correct operation.

As a first generalization of the detection data, each scanner concurrently generated a compressed log format where successive detections of the same device within 10.24 seconds of each other were compressed into detection intervals. This duration corresponds to the standard Bluetooth inquiry time (Peterson et al., 2006). The difference between the detections and the resulting detection intervals is illustrated in figure 5.2 on page 95. The resulting dataset, hereafter referred to as the raw data, consisted of 17,496 Bluetooth devices being

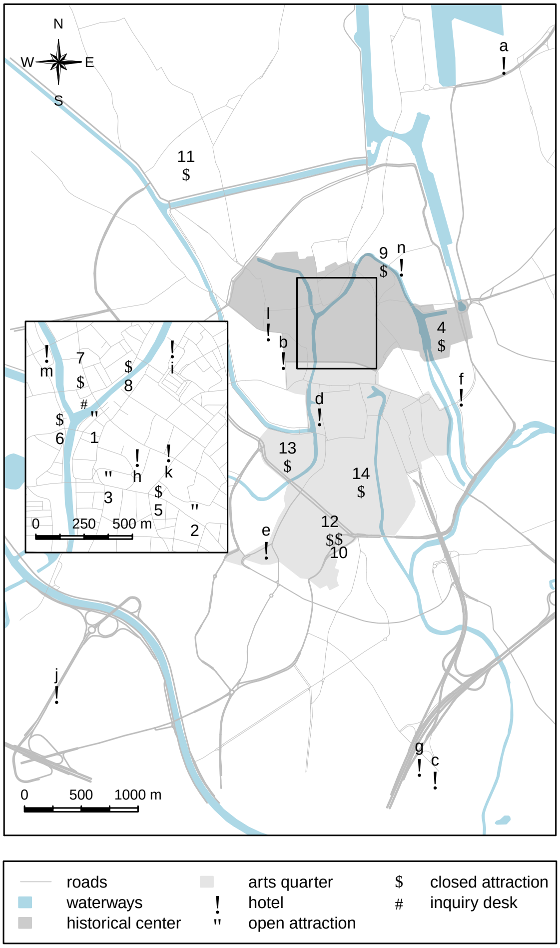


Figure 5.1: Overview of the Bluetooth sensor placement in Ghent, Belgium. The inset map shows a more detailed view of the city center.

detected over 235,597 time intervals over all locations. As tracked individuals were not approached, no additional socio-demographic or other variables were present in the dataset. The owners of the detected devices thus remain completely anonymous, and were in fact not aware of being part of a scientific study. In previous experiments, we observed that around 8% of a general public is traceable through a detectable Bluetooth device with the class 'phone'. We will use this figure to provide an approximation of the number of detected individuals based on the number of detected phones. We continue this chapter by describing association rule learning.

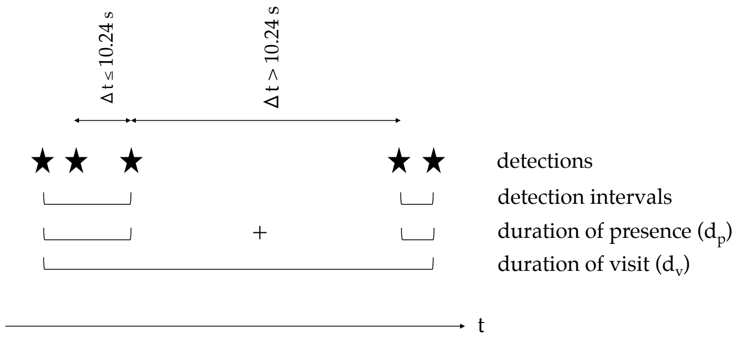


Figure 5.2: Schematical representation of detections, detection intervals, the duration of presence (d_p), and the duration of visit (d_v).

5.2.2 Association rule learning

Association rule learning represents a popular data mining method for discovering interesting relationships between variables in large databases. Adhering to the original definition (Agrawal et al., 1993), an association rule can be defined as $X \Rightarrow Y$ with $X, Y \subseteq I$ and $X \cap Y \neq \emptyset$. The itemsets X and Y are called *antecedent* and *consequent* respectively. The total itemset I in this study consists of the fourteen attractions: $I = \{1, 2, \ldots, 14\}$. The database of transactions can be formalized as $D = \{t_1, t_2, \ldots, t_n\}$ with each transaction $t_i \subseteq I$. Note that the general notion of a transaction is borrowed from the domain of market basket analysis (Chen et al., 2005), but consists of an unordered set of locations visited by a Bluetooth device i. The time ordering of visits is thus ignored, which distinguishes the method from sequential pattern mining. The rules are generated by the *Apriori* algorithm (Agrawal and Srikant, 1994) with the *arules* package (Hahsler et al., 2011) in the R (2.14.0) statistical environment.

Three measures are used to compare rules: support, confidence and lift. The *support* of a rule is a measure of the share of tracked individuals to which the rule applies: $s(X \Rightarrow Y) = s(X \cup Y)$, with the support of an itemset Z (in this case $Z = X \cup Y$) defined as the proportion of transactions in the dataset which contain the itemset: $s(Z) = {}^{\#\{t_i \in D : Z \subseteq t_i\}}/_{\#D}$. The *confidence* of a rule is a measure of the probability of its consequent given its antecedent: $c(X \Rightarrow Y) = {}^{s(X \cup Y)}/_{s(X)}$. The *lift* of a rule is a measure of its support compared with the support that can be expected if X and Y were independent: $l(X \Rightarrow Y) = {}^{s(X \cup Y)}/_{s(X)s(Y)}$. As this measure indicates whether a rule's support is lower, similar or higher than would be expected if X and Y are assumed independent, it is often used as the primary measure for the interestingness of a rule. In other words, rules with a higher lift indicate a stronger association between antecedent and consequent than what could be predicted based on the frequency of the items separately and are thus potentially more informative and valuable. In order to further clarify these concepts, table 5.1 on the following page shows an illustrative example of transactions (visitors) being constituted of visits to four of the main tourist attractions in Paris. Say, the following rule is generated: $\{Louvre\} \Rightarrow \{Arc\,de\,Triomphe, Notre\,Dame\}$. The support of this rule would be the share of visitors that visited all three attractions: $s = {}^{2}/_{5} = 0.4$. Its confidence would be its support divided by the support of its antecedent: $c = {}^{2/5}/_{3/5} \simeq 0.67$. The lift would calculated as: $l = {}^{2/5}/_{3/5 \times 3/5} \simeq 1.11$. To limit the number of generated rules and enhance interpretability, the following constraints were used for the *Apriori* algorithm: $s \geq {}^{5}/_{\#D}$, $c \geq 0.05$ and $\#X + \#Y \geq 1$ (in order to filter out rules with an empty antecedent). This way, rules need to be supported by at least 5 tracked individuals or roughly 60 individuals taking the detection ratio of 8% into account. We finish this chapter with more details on how we will present the information extracted from this data mining method.

Table 5.1: Illustrative example of a transaction database in the context of tourist attractions in Paris (1: visited, 0: not visited).

Transaction (=visitor)	Eiffel Tower	Louvre Museum	Arc de Triomphe	Notre Dame
1	1	1	0	1
2	0	0	1	0
3	1	0	1	1
4	1	1	1	1
5	0	1	1	1

5.2.3 Visit pattern maps

A correct dissemination of patterns or knowledge discovered through data mining is essential. The output of this specific case study should be tailored for all stakeholders involved in the tourism management of the study area. Association rule learning methods are known to generate large amounts of rules, and the selection of those rules with a higher relevance to the research question is a non-trivial task. Several approaches in visualizing association rules, in contrast with the classical tabular representation, have already been documented. These include the use of scatter plots and matrix-based visualizations (Hahsler and Chelluboina, 2010), graph-based representations (Appice and Buono, 2005), parallel coordinate plots (Bruzzese and Davino, 2003; Yang, 2005), 3D volumes (Compieta et al., 2007), or others (Techapichetvanich and Datta, 2005). As a way of summarizing the gathered knowledge on tourist attraction visits of a specific segment of individuals, we introduce an alternative approach called a 'visit pattern map'. This map is a geographical depiction combining two types of information: the spatial distribution of visits over the study area, and the association (combination) of visits to different attractions. The spatial distribution of visits is visualized by proportionally sized circles showing the share of tracked individuals that visited each attraction. The association between the different attractions is visualized by means of lines connecting different attractions. We believe that a geographical depiction of association rules will enhance the interpretability, in contrast with the traditional tabular fashion of representation. In order to avoid cluttering, we only visualize rules with single-item antecedents. This way, rules are only associated with one item in both the antecedent and consequent. Rules can now be represented by a single line connecting two attractions at their geographical location. Although it could be possible to include all three indicators in the visualization, we opt to neglect the confidence as it is known to be biased by frequent items in the consequent (Tan et al., 2005). Including it would additionally clutter the visualization as it is the only measure which is not symmetrical for the case of two-item rules ($s(a \Rightarrow b) = s(b \Rightarrow a)$, $l(a \Rightarrow b) = l(b \Rightarrow a)$, $c(a \Rightarrow b) \neq c(b \Rightarrow a)$). The support of a rule is linked to the width of the line, the lift is represented by a discontinuous color scale. Rules with higher lift values are plotted after (above) rules with lower lift values, making the former easier to identify. The visit pattern maps generated for this dataset are shown in figure 5.7 on page 107 and figure 5.8 on page 108.

5.3 Filtering

As mentioned above, nothing was known on the individuals carrying the detected Bluetooth devices. Before analyzing the dataset for associations between different attractions, we needed to ascertain whether devices detected at a certain attraction represented actual visitors or individuals that merely passed the location (either because of a sensor with a detection range that was too large, because these individuals physically approached a registration desk but only for information purposes, or in the case of staff). Analogously, we needed to distinguish between actual hotel guests and detected individuals with other purposes (hotel staff, restaurant guests, convention attendees, browsers, etc.). In order to make this distinction, we applied a progressive filtering process on the set of detected Bluetooth devices at each location. The filtering was based on a combination of three parameters: the type of device (accessible through the COD code), the duration of visit (d_v) to a location and the duration of presence (d_p) at a location. By taking the device type into account, we could filter on 'phones' and remove all other classes that do not represent a moving individual (car kits, printers, etc.). The duration of *visit* was calculated as the time difference between the very first and very last detection of a device at the corresponding location. In contrast, the duration of *presence* is the duration that a device was actually detected by the sensor of that location (after merging of co-located detection intervals, i.e. intervals that were less than one minute apart). This way, a device that was detected at a location from 20:35 until 20:42 and from 08:32 until 08:35 the next morning would have a duration of visit d_v of 12 hours and a duration of presence d_p of 10 minutes. The concepts of d_v and d_p are further illustrated in figure 5.2 on page 95.

For most locations, we received actual visitor/guest counts for the fifteen day tracking period (n_v). By taking the detection ratio of 8% (ρ) into account, we could estimate the number of visitors that should have been tracked at each location as such: $n'_v = n_v \times \rho/100$. Where available, we could compare these figures with the actual number of detected devices (n_d) that remained after each successive filtering step. The ratio $\delta = n_d/n'_v$ could then act as a reference for stating when the filtering process had reached an appropriate end point ($\delta \rightarrow 1$). The limits imposed on d_v and d_p were chosen by heuristic common sense linked to the type of venue. As such, we describe the filtering for the group of hotels, open attractions and inquiry desk, and closed attractions separately. Figure 5.3 represents a graphical overview of the filtering process. In the remainder of the chapter, all visitor counts will refer to detected visitors (n_d) unless otherwise stated.

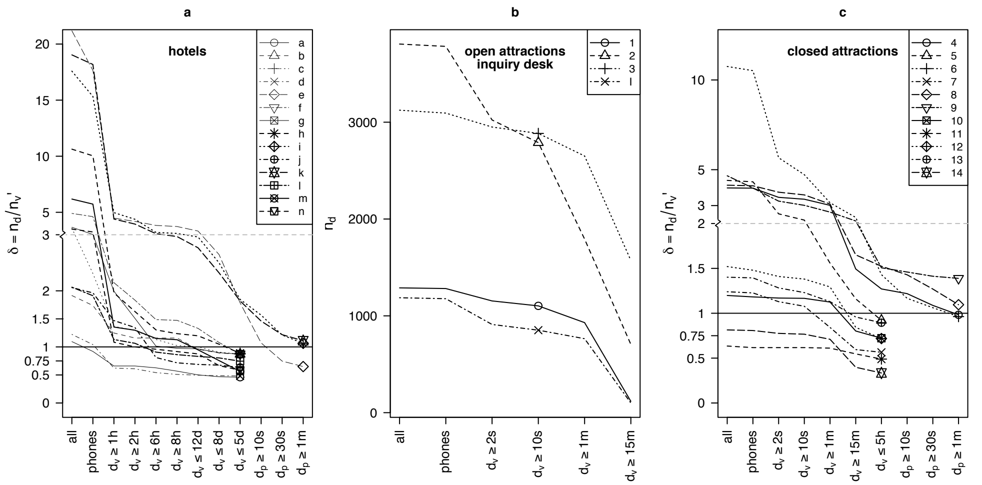


Figure 5.3: Progressive filtering process on the detected Bluetooth devices for the hotels (a), open attractions and tourist inquiry desk (b) and closed attractions (c). The point symbols indicate the filtering end points for each location. For the hotels and closed attractions, the filtering is based on $\delta = n_d/n'_v$ with n_d representing the number of devices and n'_v the estimated number of tracked visitors/guests based on a detection ratio ρ of 8%. For the open attractions and the inquiry desk, no visitor counts were available and only n_d was taken into account. Note the breaks in scale depicted as the dashed horizontal line on the y-axes in (a) and (c).

As can be seen in figure 5.3a, the number of unfiltered devices detected at most hotels significantly exceeded estimates based on guest counts (with hotels e, i and k being the extremes with $\delta \simeq 20$). Only hotels a and d seemed to represent a set of devices with an acceptable size without any filtering. A constraint on phones caused a moderate decrease in δ for all hotels, but a significantly larger decrease in hotel c. Further investigation indicated this sensor's range overlapping with a nearby parking lot, causing an overrepresentation of devices associated with vehicles (28% vs. 7 ± 4% for all other hotels). The subsequent filtering steps were based on d_v. Setting the lower limit to one hour clearly removed the largest share of noise from the dataset, but common sense dictates that a visit to a hotel should range from at least 8 hours (a guest checking in at night, and checking out early next morning), up to a reasonable maximum number of days (in order to filter out hotel staff, subcontractors, etc.). The upper limit was fixed at 5 days, which is still rather conservative compared to the average duration of a visit in Flanders of 2,43 days (Toerisme Vlaanderen, 2012). After filtering on the visit duration, all hotels except hotels e, i and k were now associated with sets of devices that corresponded with or were slightly smaller than the estimations. For these three hotels, a further filtering on the actual duration of presence yielded acceptable sets when the lower limit was set at one minute. The detection of individuals frequently passing these hotels or their registration desks over several days is the most probable reason for the necessity of this extra filtering step. We opted to apply this additional filtering solely on these three hotels, because it caused a significant additional decrease in four other hotels.

For the three open attractions and the tourist inquiry desk (figure 5.3b), no visitor counts were available and the filtering was based on a conservative minimum visit duration of 10 seconds. This choice may seem arbitrary, but was made on the notion that some of these attractions are known to serve as passageways for general movements throughout the center. A further distinction between individuals merely glancing at the attraction, and purposeful visits will not only necessitate further data but also entails a semantic discussion on how to define a 'visit' to such a location.

The filtering for the closed attractions was again based on a combination of the constraint to phones, duration of visit, and the duration of presence. The heuristic lower and upper limits of the visit duration were now set to 15 minutes and 5 hours respectively, thus filtering out inquirers and museum staff. Figure 5.3c shows that all but four attractions reached an acceptable δ value after filtering on d_v alone. As with the hotels, a further filtering on the duration of presence (minimum of 1 minute) was necessary for the remaining venues. Only attraction 9 was associated with a device set that is still somewhat larger than would be expected after this filtering ($\delta = 1.4$). In absolute numbers, the difference seems less pronounced ($n_d = 59$ vs. $n'_v = 42$). As the lower limit on d_p would need to be set at nearly 5 minutes, we chose to stop the filtering and accept one device set that was slightly larger than expected. The nearby presence of a bar associated with the museum (the visitors of which could be tracked but are not included in the visitor counts) might have caused this anomaly.

5.4 Data exploration

As a summary of the progressive filtering process, the size of each filtered device set associated with all covered locations is depicted in figure 5.4 on page 101. The filtered sets can now be aggregated into three different sets of tracked individuals: visitors (symbolized as V, part of at least one of the filtered device sets at attractions 1–14), hotel guests (symbolized as H, part of at least one of the filtered device sets at hotels a–n) and inquirers (symbolized as I, part of the filtered device set at the tourist inquiry desk). The total filtered population of tracked individuals $P = H \cup V \cup I$ contains 7,326 devices, which represents a 58% reduction by the filtering process. Looking at the number of hotel guests and comparing with the map of the study area depicted in figure 5.1 on page 94, we can generally distinguish between cheaper hotels further from the center with lower guest numbers, larger and more expensive hotels in the center, and two hostels accommodating a very small share of guests. The open attractions are associated with significantly larger numbers of visitors than the closed attractions. Based on this finding and the previously mentioned different characteristics of a visit/visitor between the two types of attractions, we will further distinguish between visitors *sensu lato* ($V = V_o \cup V_c$) and visitors *sensu stricto* (V_c), with V_o representing all visitors to at least one of the open attractions and V_c all visitors to at least one of the closed attractions.

As already stated in section 5.3 on page 97, there is a significant semantic difference between visiting an open and a closed attraction. Whereas a visit to an open attraction can be very short and coincidental in nature due to the free entrance, a visit to a closed attraction represents a significantly longer and probably more deliberate choice. We therefore suspect that both types of visits are generally performed by different individuals. In order to explore this hypothesis, we start by defining 5 visitor segments based on (different combinations of) the sets V, V_o and V_c. These five segments are: V (visited at least one open/closed attraction), V_o (visited at least one open attraction), V_c (visited at least one closed attraction), $V_o \setminus V_c$ (visited at least one open attraction but none of the closed attractions), $V_c \setminus V_o$ (visited at least one closed attraction but none of the open attractions). The similarities between these segments are depicted in table 5.2 on page 102, where Jaccard indices (size of the intersection divided by size of the union) were calculated as measures for the similarity. As expected, the number of open attraction visitors V_o (80% of V) clearly exceeds the number of closed attraction visitors V_c (36%). Additionally, the overlap is quite small: 80% of the open attraction visitors never visited any of the closed attractions, 56% of the closed attraction visitors never visited any of the open attractions, only 16% combined both types of attractions. Next, we investigated the share of hotel guests, inquirers and one-day/several-day visitors (calendar days) for the different visitor segments. It appears that especially individuals that only visited one or more closed attractions ($V_c \setminus V_o$) show a deviating (lower) hotel (4 vs. 8%) and inquiry desk (4 vs. 13%) use, and contain a slightly higher frequency of one-day visitors (83 vs. 78%) compared to V. It would be reasonable to assume that this is caused by the geographical distance between most of the closed attractions and the historical center where most tourists (identified hotel guests) stay, making that this group has a higher representation of local (one day) visitors.

Almost two thirds of all visitors (V) only visited one or more of the open attractions ($V_o \setminus V_c$). Due to the low number of open attractions (3), this would make a large part of the dataset be composed of a rather homogeneous and less informative subpopulation. As such, we continue exploring the set V_c (i.e. the individuals having made at least one deliberate choice in registering for and visiting one of the closed attractions) as a more heterogeneous and better candidate set for a data mining method. We continue the investigation by distin-

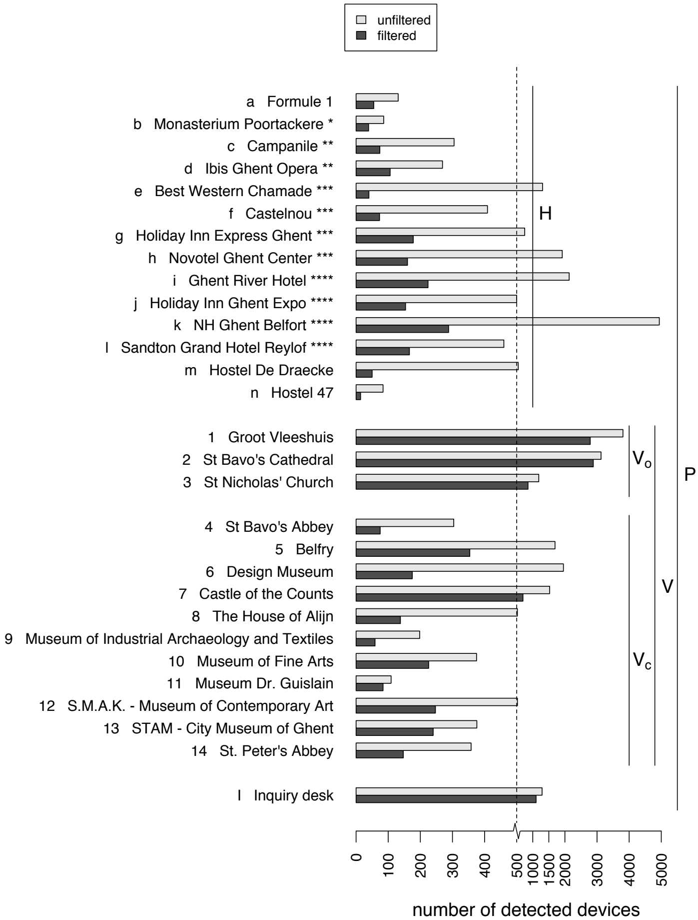


Figure 5.4: Preprocessing summary showing the number of detected devices before and after filtering at each location, and the aggregation of the filtered devices into the sets of hotel guests H, visitors V (*sensu lato*), information seekers I, open attraction visitors V_o, closed attraction visitors V_c, and the entire population of tracked individuals P. Note the break in scale on the x-axis depicted by the dashed vertical line.

guishing between one-day and several-day visitors, leading to two extra visitor segments: $V_c \cap P_{1d}$ and $V_c \cap P_{>1d}$ respectively. As table 5.2 on the next page shows, one in four visitors in V_c is present in the database over more than one (calendar) day. Around 37% of the several-day visitors were identified as hotel guests, and 23% went to the inquiry desk. Please note that the hotel usage will be an underestimation of the real figure as only 67% of the total hotel capacity was tracked in the experiment, so the other 63% of several-day visitors will be a combination of hotel guests outside of the tracked sample and visitors that performed visits over several days that may or may not be contiguous (e.g. over a first visit

Table 5.2: Sizes of, similarities between different visitor segments (Jaccard index), and the corresponding share of hotel guests (H), inquirers (I), one-day (P_{1d}) and several-day visitors ($P_{>1d}$) for each visitor segment. The five visitor segments are: V (all visitors), V_o (open attraction visitors), V_c (closed attraction visitors), $V_o \setminus V_c$ (only visited one or more open attractions but none of the closed attractions), $V_c \setminus V_o$ (only visited one or more closed attractions but none of the open attractions).

		Jaccard index: $\#(A \cap B)/\#(A \cup B)$					$\#(A \cap B)/\#A$						
$A \downarrow B \rightarrow$	$\#A$	V	V_o	V_c	$V_o \setminus V_c$	$V_c \setminus V_o$	H	I	P_{1d}	$P_{>1d}$	V	V_o	V_c
V	5,891	1	0.80	0.36	0.64	0.20	0.08	0.13	0.78	0.22			
V_o	4,726	0.80	1	0.16	0.80	0	0.09	0.15	0.77	0.23			
V_c	2,095	0.36	0.16	1	0	0.56	0.10	0.14	0.75	0.25			
$V_o \setminus V_c$	3,796	0.64	0.80	0	1	0	0.07	0.12	0.80	0.20			
$V_c \setminus V_o$	1,165	0.20	0	0.56	0	1	0.04	0.04	0.83	0.17			
$V_c \cap P_{1d}$	1,564						0	0.11	1	0			0.75
$V_c \cap P_{>1d}$	531						0.37	0.23	0	1			0.25
$V_c \cap P_{>1d} \cap H$	196						1	0.27	0	1			0.09
H	1,581							0.08			0.31	0.28	0.13
H_{far}	456							0.05			0.16	0.15	0.03
H_{4*}	675							0.09			0.37	0.33	0.16
H_{hostel}	64							0.02			0.39	0.33	0.19

in the first weekend, and a second visit in the following weekend). As a last segment, we additionally constrained to identified hotel guests ($V_c \cap P_{>1d} \cap H$). This segment, which comes as close to the definition of tourists as explained in the introduction, shows an even slightly larger use of the inquiry desk.

In order to explore the possible effect of hotel choice on visiting patterns, we further also distinguish between the guests of hotels located far from the center (a, c, g or j; H_{far}), four-star hotels ($i - l$, H_{4*}) and hostels (m, n; H_{hostel}). Guests of a hotel far from the center clearly less often visit an open or closed attraction, or the inquiry desk. The four-star hotel guests seem to follow the pattern of the more general set of hotel guests. Hostel guests hardly visit the inquiry desk, but seem to visit tourist attractions slightly more often than the average hotel guest.

We also investigated the total duration covered in the tracking data and the number of visited attractions in the different visitor segments. Figure 5.5 on the facing page shows the resulting distributions of the number of calendar days, number of attractions and number of closed attractions for the ten visitor segments which will be further investigated in section 5.5 on page 105. Around 80% of the visitor population V is tracked over only one calendar day, and practically none over more than five days. Visitors to at least one closed attraction (V_c) and those that did not visit any open attraction ($V_c \setminus V_o$) do not seem to deviate significantly from this distribution. Several-day visitors that were identified as hotel guests seem to cover a slightly higher number of calendar days than those that were not. Please recall that we cannot state with certainty whether an individual did not stay at a hotel because only a subset of hotels was covered by a Bluetooth sensor, so it is difficult to explain this difference directly. Concerning the hotel-based visitor segments, guests of the remote hotels cover slightly less calendar days, and hostel guests slightly more than on average. More than 60% of the visitors *sensu lato* only visit one attraction. Those that visited at least one closed attraction are more distributed towards a higher number of attractions. Looking at the number of visited closed attractions, however, the share of visitors that only visited one closed attraction is even higher (over 80%). Several-day visitors (both those identified as hotel guests and those that were not) visit a larger number of attractions when the open attractions are included, but do not deviate significantly from the general trend that most visitors only visited one closed attraction. Remote hotel guests visit slightly less closed attractions, hostel guests slightly more.

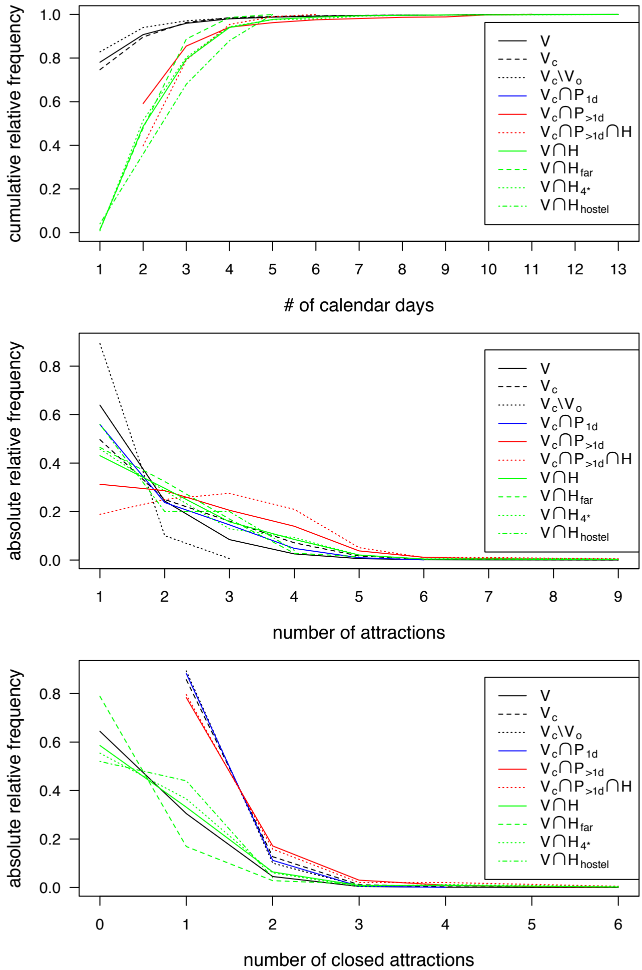


Figure 5.5: Cumulative relative frequency of the number of tracked calendar days (top), and absolute relative frequency of the number of visited attractions (middle) and *closed* attractions (bottom).

As a first approach to investigating the degree of association between the different venues, we also examine the degree of overlap in between the sets of visitors/guests of the different locations as listed in figure 5.4 on page 101. These overlaps are again calculated by the Jaccard index and are depicted in figure 5.6 for both the unfiltered and filtered sets. As expected, the overlaps between the hotels is very low: tourists usually only stay at one hotel during a visit. The remaining overlaps after filtering are mainly caused by geographical proximity and guests of one hotel thus being mistakenly classified as guests of the other (e.g. hotels h and k nearly face each other across the same street) . The open attractions, and closed attractions 5 ('Belfry') and 7 ('Castle of the Counts') all show significant mutual overlaps. As can be seen in figure 5.1 on page 94, proximity is probably the most important cause besides similarity in characteristics (all are historical buildings). The same effect also explains the higher overlaps between the tourist inquiry desk and the open attractions in the center, and the moderate overlaps with attractions 5–7. Outside of the center, attractions 10, 12 and 14 in the arts quarter also show significant mutual overlaps. A deeper understanding of the associations, however, will be mined for by the association rule learning method. In the next section, we will outline its results.

unfiltered — hotels

	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a		0	0	1	0	0	0	0	0	0	0	0	0	1
b	0		1	0	0	0	0	0	0	0	0	0	0	0
c	0	1		0	0	0	4	0	0	0	0	1	0	0
d	1	0	0		1	0	0	1	1	0	1	1	1	0
e	0	0	0	1		1	1	2	2	0	2	1	1	0
f	0	0	0	0	1		1	1	1	0	1	1	1	1
g	0	0	4	0	1	1		1	1	2	1	1	0	0
h	0	0	0	1	2	1	1		3	0	17	2	2	0
i	0	0	0	1	2	1	1	3		0	5	1	2	1
j	0	0	0	0	0	0	2	0	0		1	0	0	0
k	0	0	0	1	2	1	1	17	5	1		2	2	0
l	0	0	1	1	1	1	1	2	1	0	2		2	0
m	0	0	0	1	1	1	0	2	2	0	2	2		1
n	1	0	0	0	0	1	0	0	1	0	0	0	1	

filtered — hotels

	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a		0	0	1	0	0	0	0	0	0	0	0	0	0
b	0		0	0	0	0	0	0	0	0	0	0	0	0
c	0	0		0	0	0	2	0	0	0	0	0	0	0
d	1	0	0		0	0	0	0	0	0	0	0	0	0
e	0	0	0	0		0	0	1	0	0	1	0	0	0
f	0	0	0	0	0		0	0	0	0	0	0	0	0
g	0	0	2	0	0	0		0	0	0	0	0	0	0
h	0	0	0	0	1	0	0		0	0	4	0	0	0
i	0	0	0	0	0	0	0	0		0	0	0	0	1
j	0	0	0	0	0	0	0	0	0		0	0	0	0
k	0	0	0	0	1	0	0	4	0	0		0	0	1
l	0	0	0	0	0	0	0	0	0	0	0		0	0
m	0	0	0	0	0	0	0	0	0	0	0	0		0
n	0	0	0	0	0	0	0	0	1	0	1	0	0	

unfiltered — open and closed attractions, inquiry desk

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	I
1		24	12	1	13	14	16	5	1	1	0	2	2	1	14
2	24		22	1	22	14	22	5	1	2	0	1	1	1	14
3	12	22		1	17	10	16	5	1	2	0	1	1	1	12
4	1	1	1		1	1	1	4	1	2	0	6	1	6	1
5	13	22	17	1		11	14	4	1	1	0	1	1	1	10
6	14	14	10	1	11		12	5	1	2	0	2	1	2	13
7	16	22	16	1	14	12		6	1	2	0	1	1	1	15
8	5	5	5	4	4	5	6		2	2	0	2	1	2	8
9	1	1	1	1	1	1	1	2		2	0	2	1	1	1
10	1	2	2	2	1	2	2	2	2		1	21	2	11	2
11	0	0	0	0	0	0	0	0	0	1		0	1	1	0
12	2	1	1	6	1	2	1	2	2	21	0		3	20	2
13	2	1	1	1	1	1	1	1	1	2	1	3		3	2
14	1	1	1	6	1	2	1	2	1	11	1	20	3		1
I	14	14	12	1	10	13	15	8	1	2	0	2	2	1	

filtered — open and closed attractions, inquiry desk

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	I
1		25	11	0	5	2	9	2	0	1	0	1	1	0	14
2	25		19	0	8	3	12	2	0	2	0	1	1	0	14
3	11	19		0	10	2	9	2	1	2	0	1	1	1	12
4	0	0	0		0	0	0	2	2	0	0	1	1	0	1
5	5	8	10	0		3	9	3	1	1	0	1	0	1	6
6	2	3	2	0	3		3	2	0	2	0	1	1	1	4
7	9	12	9	0	9	3		4	1	1	0	1	1	1	9
8	2	2	2	2	3	2	4		3	1	1	1	1	1	2
9	0	0	1	2	1	0	1	3		2	0	0	1	0	1
10	1	2	2	0	1	2	1	1	2		1	8	2	3	2
11	0	0	0	0	0	0	0	1	0	1		0	1	0	0
12	1	1	1	1	1	1	1	1	0	8	0		1	10	1
13	1	1	1	1	0	1	1	1	1	2	1	1		2	1
14	0	0	1	0	1	1	1	1	0	3	0	10	2		1
I	14	14	12	1	6	4	9	2	1	2	0	1	1	1	

Figure 5.6: Overlap between unfiltered (left) and filtered (right) device sets at hotels (a–n, top), and the open and closed attractions and inquiry desk (1–14 + I, bottom). The numbers in the grid are the Jaccard indices of each combination of device sets, and represent the degree of overlap (0: less than one % overlap, 100: completely identical).

5.5 Visit pattern mining

The degree of overlap between the different locations and their visitors discussed in the previous section offers a first insight into the degree of association between the different attractions. To obtain a deeper understanding, however, a more thorough analysis is needed. In this section, we will mine for association rules between the fourteen (open and closed) attractions for the previously identified visitor segments. The mining process was described in section 5.2.2 on page 95 (minimum support of 5 devices, minimum confidence of 5%). The subset of rules with only one item in the antecedent is visualized in their geographical context, together with the share of visitors in the segment that visited each attraction, in a 'visit pattern map' as described in section 5.2.3 on page 96. For each segment, the top-20 of all association rules is additionally listed in a tabular fashion. The rest of this section is divided into two parts. In section 5.5.1 we describe the patterns and their differences found going from the general set of visitors *sensu lato* V to the most specific segment of hotel guests that visited at least one closed attraction and have a duration of visit of at least 1 day ($V_c \cap P_{>1d} \cap H$). In section 5.5.2 on the following page, we investigate the potential differences in patterns for remote hotel guests, four-star hotel guests, and hostel guests.

5.5.1 Visitor segment exploration

Figure 5.7 on page 107 shows the visit pattern maps for the five visitor segments going from visitors *sensu lato* V, over V_c, $V_c \cap P_{1d}$, and $V_c \cap P_{>1d}$ to $V_c \cap P_{>1d} \cap H$. Table 5.3 on page 113 lists the top-20 of all rules (including those with more than one item in the antecedent) for all these visitor segments. Looking at the share of visitors visiting each attraction (proportionally-sized circles) on the visit pattern map for visitors *sensu lato* (V), we clearly observe a concentration of visits in the city center and its open attractions (attractions 1 and 2 each attract nearly 50% of the tracked population). Of the closed attractions, the 'Castle of the Counts' (7) attracts the largest share of visitors (12%), followed by the 'Belfry' (5) with 6%. All other closed attractions, both in the center and more remote, attract significantly smaller shares — the largest being the museum of contemporary art ('SMAK', 12) which attracts around 4% of the population. Concerning the rules for the visitors *sensu lato* (V) segment, both the map and the top-20 show that all association rules with a high lift have a very low support (the rule with the highest lift in the map is supported by only 1% of the population, rules with more items in the top-20 have even lower supports). As previously mentioned, visitors *sensu lato* represent a heterogeneous group of individuals — many of which only visited one or more open attractions (figure 5.5 on page 103) on relatively short trips. Despite low supports, the strong associations between the 'SMAK' museum (12) and the museum of fine arts (10) and especially between the 'SMAK' museum and the 'Saint Peter's abbey' (14) are noteworthy. Longer-distance associations have a low lift and support, indicating that visitors seem to rarely combine in-center and out-of-center visits. The associations between the open attractions logically have a larger support, but show lift values close to 1 suggesting that the associations are not significantly stronger than expected.

With regard to the more specific visitor segments and starting by constraining to closed attraction visitors (V_c), the associations between the three art museums in the south are no longer the strongest in the set (although they still have lift values higher than 1). Instead, the highest lift values are now found in the city center, especially between 'Saint Nicholas' Church' (3), and the 'Belfry' (5) and 'Saint Bavo's Cathedral' (2): for visitors that made at least one conscious choice in visiting a closed attraction we find a higher than expected association between the attractions in the city center (lift>1).

Visitors *sensu stricto* that were only tracked on one calendar day ($V_c \cap P_{1d}$) show nearly no association between attractions in the center and those outside of the center, revealing they either visit one or the other.

Those that were tracked over several days ($V_c \cap P_{>1d}$) do show these associations as they probably have more free time to cover these distances. Comparing these visitors to the visitors *sensu stricto* in general (V_c), there are no clear differences in both patterns for the city center.

Finally, comparing identified hotel guests ($V_c \cap P_{>1d} \cap H$) with the general several day visitors ($V_c \cap P_{>1d}$) some differences between the patterns become visible. 'Saint Peter's abbey' looses its significance, together with the associations between the art museums in the south. In contrast, both the 'House of Alijn' (8) and the city museum of Ghent ('STAM', 13) appear to gain in importance: both now show significant (lift>1) and considerable (support of almost 10%) associations. The lift of the association between the Museum of Industrial Archeology and Textiles ('MIAT', 9) and attraction 1 also seems to have risen, but the support of the rule is very low (3,5%, 7 tracked individuals).

5.5.2 Effect of hotel location and type

The potential effect of the hotel location and type is investigated by comparing the visit patterns of four distinct segments of hotel guests: the general set of hotel guests that visited at least one open/closed attraction ($V \cap H$), those that stayed at one of the four most remote hotels (a, c, g or j; $V \cap H_{far}$), those that stayed at one of the four 4-star hotels ($V \cap H_{4*}$), and finally those that stayed in one of both hostels ($V \cap H_{hostel}$). The four corresponding visit pattern maps are shown in Figure 5.8 on page 108. The top-20 of rules for each segment is listed in table 5.4 on page 114. Remote hotel guests, who were already shown to significantly more often engage in one or more visits (table 5.2 on page 102), clearly show a preference for the open attractions and the 'Castle of the Counts' (7) in the center. They rarely visit any of the closed attractions further from the center, and associations besides those between open attractions are rare. The 4-star hotel guests show a visit pattern that is almost identical to the general visit pattern for hotel guests, which makes sense as they form the largest share of hotel guests. Hostel guests, finally, show a pattern that bears some similarities to the remote hotel guests. The museum of contemporary art ('SMAK', 12) seems to attract a significantly

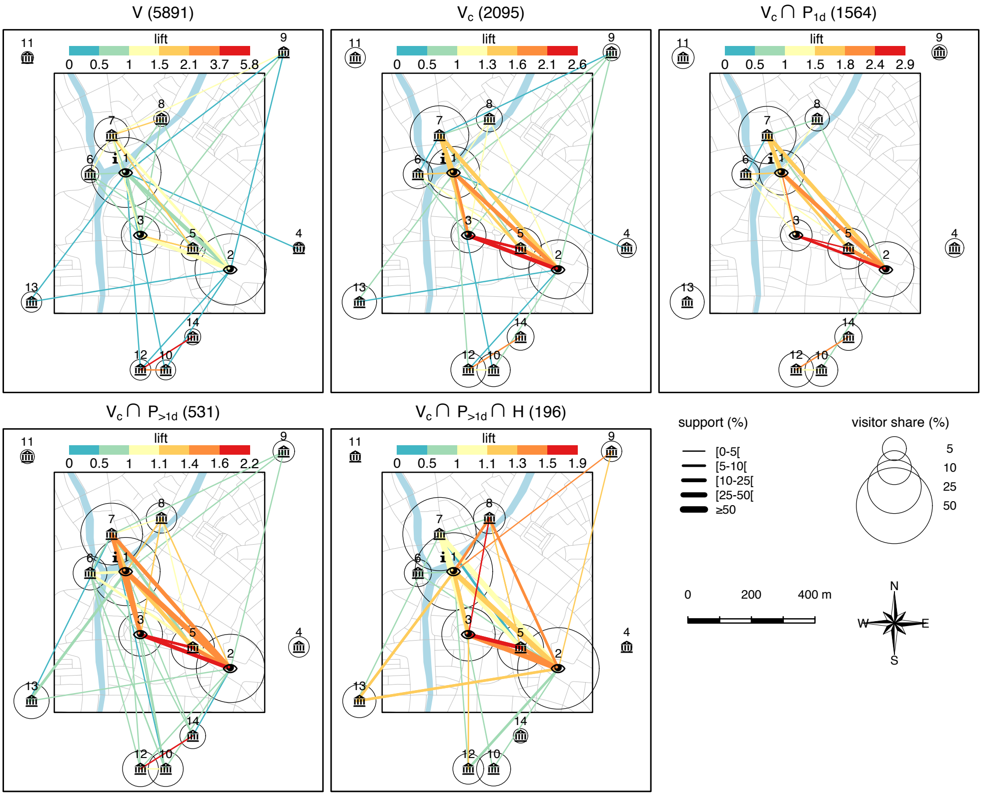


Figure 5.7: Visit pattern maps for visitor segments V, V_c, $V_c \cap P_{1d}$, $V_c \cap P_{>1d}$, and $V_c \cap P_{>1d} \cap H$. The size of each segment is given between brackets. The spatial distribution of visits is represented by proportionally sized circles symbolizing the share of visitors in the segment visiting each attraction. Two-element association rules are visualized as lines connecting the locations in the antecedent and consequent. The support of a rule is symbolized by the width of the line, the lift by its color (as shown by the bars above each map, classification was equal-range for lifts below 1 and according to Jenks natural breaks above 1). Attractions far from the city center (depicted by the rectangle) are not depicted on their actual geographical location in order to increase the legibility of the visualization.

larger share of these visitors (24% vs. 4% of the general hotel guest segment). It should be noted, however, that the size of this last segment has become rather small, possibly limiting the representativeness of the pattern it exhibits.

5.6 Discussion and conclusion

In this section, we will first further interpret the filtering and mining processes and the results that were generated from the dataset. Subsequently, we will discuss the current and future potential of the presented methodology (Bluetooth tracking + association rule

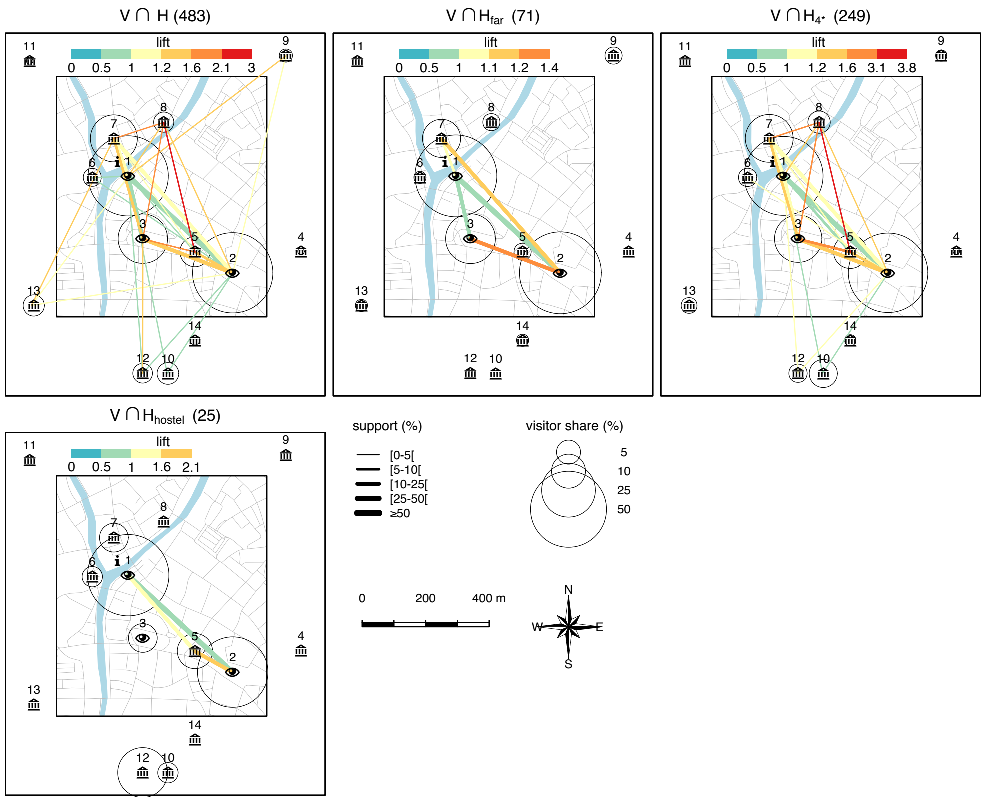


Figure 5.8: Visit pattern maps for visitor segments $V \cap H$, $V \cap H_{far}$, $V \cap H_{4*}$, and $V \cap H_{hostel}$. The size of each segment is given between brackets. The spatial distribution of visits is represented by proportionally sized circles symbolizing the share of visitors in the segment visiting each attraction. Two-element association rules are visualized as lines connecting the locations in the antecedent and consequent. The support of a rule is symbolized by the width of the line, the lift by its color (as shown by the bars above each map, classification was equal-range for lifts below 1 and according to Jenks natural breaks above 1). Attractions far from the city center (depicted by the rectangle) are not depicted on their actual geographical location in order to increase the legibility of the visualization.

learning) for tourism management purposes. We conclude with further avenues for future research.

5.6.1 Further interpretation of filtering, mining and results

As in any knowledge discovery process, data mining techniques form only part of a chain of subprocesses going from data to knowledge (Fayyad et al., 1996). As described in section 5.3 on page 97, the Bluetooth tracking data needed to undergo significant preprocessing before being able to function as an input for association rule learning. There are two main reasons for the need for such an extensive filtering procedure. First, Bluetooth is a very popular and widely distributed technology available on a large variety of devices. Whereas phones can be assumed to be linked to one individual each, other classes, such as carkits, cannot. Second, the proximity of an individual to a place of interest cannot be directly associated with a certain activity related to that location. This becomes clear in figure 5.4 on page 101, where most sensors at both hotels and attractions clearly detected significantly more Bluetooth devices than predicted according to the actual guest/visitor counts and the detection ratio of 8%. As can be seen in figure 5.3 on page 98, most hotels are associated with very short detections indicated by the large decrease in filtered devices by choosing d_v to a minimum of one hour. A further constraint of this parameter to the heuristic interval between 8 hours and 5 days leads to acceptable filtered sets, except for three hotels where a further constraint had to be placed on d_p. A similar trend, but with different constraints on d_v, is visible for the closed attractions. While it is impossible to statistically verify the accuracy of the applied temporal filtering, figure 5.6 on page 104 at least gives an indication that most significant overlaps between hotels (where there should not be any) have decreased significantly. The overlaps that remain after filtering are mainly due to proximity (e.g. hotels h and k). The uncertainty over whether a tracked individual actually visited an attraction and/or stayed at a hotel could be addressed in future work by making the scanning of the device part of the registration process.

Before we started mining for association rules, we performed an extensive data exploration. Two hypotheses received special attention. First, the combination of open and closed attractions was suspected to result in a heterogeneous group of 'visitors' because of the semantic difference between both types. Second, we needed to investigate the tendency of individuals to visit more than one attraction over the fifteen day tracking period before looking at the specific associations. As was listed in table 5.2 on page 102, visitors *sensu lato* show little overlap between visitors *sensu stricto*: only 36% of the former group also visit at least one closed attraction, and only 16% combine both types. Figure 5.5 on page 103 showed that only a small fraction of visitors combined more than one closed attraction, which would certainly influence the mining for association rules. Readers should bear in mind that fifteen days is quite a short period for capturing combination preferences between museums for visits that are not part of a chained trip of visits (e.g. for local residents), and that the tracking period would ideally be composed of the entire touristic season. Where the mined associations for these individuals would certainly be an underestimation and thus possibly not representative, identified hotel guests (tourists) can be assumed to perform visits that form part of a trip. As such, this group ($V_c \cap P_{>1d} \cap H$) received special attention in the analyses.

Subsequently, the associations between the different open and closed attractions were investigated through the application of the *Apriori* algorithm and the interpretation and visualization of the extracted rules. Combining the exploratory findings, we gradually restricted the set of visitors *sensu lato* to individuals that visited at least one closed attraction (as this indicates at least one conscious choice in their itinerary) and additionally to identified hotel guests. Figure 5.7 on page 107 shows the resulting series of visit pattern maps. Taking all visitors *sensu lato* as input, large but predictable associations in the center appear next to very small but more interesting associations between the art-oriented attractions in the 'arts quarter' of the study area. Constricting to visitors *sensu stricto* decreased this conceptual difference due to the lower frequency (higher lift) of the center based attractions and the higher frequency (lower lift) of the art attractions. Several-day visitors clearly show associations (though most of them are quite small) between the center and more remote attractions, whereas single day visitors rarely combine both areas. The constrained set of identified hotel guests, which could be argued to be the only representative pattern in the series, shows some interesting differences with the general several day visitor set (which also includes individuals that did not stay overnight). The strong associations between the art museums disappear together with the significance of 'Saint Peter's abbey' individually, but two new locations appear in considerable associations with the open attractions in the center: the city museum of Ghent ('STAM'), and the 'House of Alijn'. Subsequently, we investigated the additional differences in visit patterns between remote hotel guests, four-star hotel guests and hostel guests. Remote hotel guests show a very large preference in the open attractions in the center and rarely venture further away, whereas hostel guests show a similar pattern with the exception of a higher representation of the museum of contemporary art ('SMAK'). Four-star hotel guests followed the general pattern.

5.6.2 Potential of the employed methodology for tourism management

The potential of Bluetooth tracking in tourism management practices was illustrated by applying an association rule learning scheme on the visits to different attractions in an urban environment. In this specific case study, we focused on discovering interesting associations as indicated by attractions appearing together more often than predicted by an independent choice model. Such associations or the lack thereof can be used for a wide array of purposes. Focusing on the attractions, existing associations could either be strengthened or non-existent or weak associations could be created by applying specific promotional advertisements at each attraction, thereby urging visitors to visit other attractions as well. In contrast, the focus could also lie on the tourist and discovered patterns could be used in recommendation systems based on collaborative filtering. These recommendations could be disseminated through the use of smartphones. In a long-term strategic context, pattern maps such as those in figure 5.7 on page 107 could also be used by urban planners for optimizing tourist accessibility and facilities. The three art museums in the south of the study area (promoted as the 'arts quarter') can serve as an illustrative example. While they do seem associated with each other mutually, our analysis has also pointed out that there is little association with the attractions in the center. Planners could tackle this issue by developing the necessary tourist facilities (e.g. hotels), improving the (visibility of) public transport options between both areas, or by designing a corridor to minimize the perceived distance between the center and the 'arts quarter' (e.g. by improving pedestrian accessibility, creating more green and open spaces, etc.). The effectiveness of certain actions could also be investigated by tracking during a period both before and after the action was taken.

5.6.3 Further issues surrounding the methodology

Three main issues can be identified with regard to the used methodology: (i) errors introduced when using presence detections to deduce activities, (ii) potential bias in the tracked sample introduced by using Bluetooth technology, and (iii) lack of any metadata on tracked visitors due to the non-participatory nature of the methodology. We will shortly reflect on all of these issues.

Deducing activities from presence detections through sensors is a process which can essentially be affected by two types of errors. The first error could be labeled as false presence detections, where a sensor detects devices that do not physically enter a building or approach a desk. Filtering procedures were able to remove most of this sensor noise but further efforts are certainly warranted in order to calibrate the sensors thereby minimizing false presence detections. The second type of error is due to actual presences not automatically implying certain activities. In the context of this study, these could be caused by staff members or inquirers. Filtering on common-sense thresholds for the duration of visit and duration of presence, we were able to deduce acceptable device sets. A higher accuracy, however, will somehow imply contacting individuals and registering the MAC address of their device(s).

Second, it is possible that certain age segments, or one gender might have a higher usage rate of devices with a discoverable Bluetooth interface. Tourists might, for example, preferentially opt to turn off their mobile phones in order not to be disturbed and thus be under-represented. Despite the undeniable importance of these potential influencing factors on any tourism management incentive, it falls outside of the scope of this research to fully investigate this issue. As in any other tracking study, however, it is of vital importance that more attention should be devoted to this question.

The lack of metadata, finally, can be interpreted as the downside of using a non-participatory methodology. Instead of dealing participatory and non-participatory methodologies as complete opposites, we argue that both methodologies might be combined. Alternatively, ad-hoc sensing networks could be made semi-participatory by approaching and interviewing part of the tracked population. In our scenario, for example, individuals could be contacted in the hotel where they are staying. Again, this fell outside of the current scope but could aid in further strengthening assumptions made about visitors.

5.6.4 Future research

While we applied an association rule learning technique in this case study, other data mining tasks could be used or combined for answering other or similar research questions as those put forward in this chapter. If the order of visits were important for example, sequential pattern mining techniques could be used. Clustering methods could segment tourists based on their associations of visits. In addition, a second and longer period of tracking could serve as a point of comparison with the patterns found during the fifteen day tracking period of this study. The visit pattern maps summarize a considerable amount of information in one map, which possibly makes them challenging to interpret. User studies could clarify the way in which people read the map (e.g. by eye-tracking procedures), and the provided insights could be used to further fine-tune the visualization. The selective visualization of 2-item-rules causes some of the (potentially valuable) information contained within rules with more items to be lost. How to efficiently visualize these complex rules in a legible way is saved for future work.

Table 5.3: Top-20 (where applicable) of association rules $X \Rightarrow Y$ for visitor segments V, V_c, $V_c \cap P_{1d}$, $V_c \cap P_{>1d}$, and $V_c \cap P_{1d} \cap H$. Rules were generated based on constraints on both support ($s_a \geq 5$) and confidence ($c \geq 0.05$), and are sorted on their lift (l) value. The relative (s_r) and absolute (s_a) support, as well as the confidence (c) are also shown. Rules with only item in the antecedent (marked in bold) are also visible in figure 5.7 on page 107.

	V (5,891 devices)						V_c(2,095 devices)						$V_c \cap P_{1d}$ (1,564 devices)					
	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l
1	3,7,8	5	8.5E-04	5	0.63	10.40	2,5,7,8	3	2.4E-03	5	0.71	6.01	12,2	10	3.2E-03	5	0.50	4.80
2	2,3,7,8	5	8.5E-04	5	0.63	10.40	2,5,8	3	3.3E-03	7	0.64	5.35	1,5,7	3	5.8E-03	9	0.39	4.05
3	10,2,7	5	8.5E-04	5	0.56	9.25	10,5	3	2.4E-03	5	0.63	5.26	5,6	1	3.8E-03	6	0.75	3.83
4	10,7	5	8.5E-04	5	0.45	7.56	10,2,5	3	2.4E-03	5	0.63	5.26	1,2,5,7	3	5.1E-03	8	0.36	3.77
5	1,2,7,8	5	8.5E-04	5	0.45	7.56	1,2,5,8	3	2.4E-03	5	0.63	5.26	2,5,6	1	3.2E-03	5	0.71	3.65
6	12,2	10	1.4E-03	8	0.29	7.45	5,7,8	3	2.4E-03	5	0.56	4.67	10,2	3	5.8E-03	9	0.35	3.59
7	2,3,8	5	1.2E-03	7	0.44	7.28	1,5,8	3	2.4E-03	5	0.56	4.67	6,7	1	5.8E-03	9	0.69	3.54
8	1,7,8	5	1.0E-03	6	0.43	7.13	10,2,7	3	2.4E-03	5	0.56	4.67	1,2,5	3	1.5E-02	24	0.34	3.50
9	2,7,8	5	1.2E-03	7	0.41	6.85	1,2,7,8	3	2.9E-03	6	0.55	4.59	2,5	3	4.0E-02	62	0.34	3.49
10	3,8	5	1.2E-03	7	0.39	6.47	5,8	3	3.3E-03	7	0.54	4.53	2,6,7	1	3.8E-03	6	0.67	3.41
11	2,3,5,7	8	8.5E-04	5	0.15	6.47	1,2,5,7	3	1.1E-02	22	0.52	4.41	1,3,6	2	3.2E-03	5	1.00	3.37
12	1,3,8	5	8.5E-04	5	0.38	6.40	1,5,7	3	1.1E-02	23	0.52	4.40	1,5	3	1.8E-02	28	0.32	3.33
13	1,2,3,8	5	8.5E-04	5	0.38	6.40	2,7,8	3	3.8E-03	8	0.47	3.96	5,7	3	1.1E-02	17	0.31	3.26
14	3,5,8	7	8.5E-04	5	0.71	6.05	10,7	3	2.4E-03	5	0.45	3.82	5,7	2	3.3E-02	52	0.96	3.25
15	2,3,5,8	7	8.5E-04	5	0.71	6.05	3,7,8	5	2.4E-03	5	0.63	3.70	1,5,7	2	1.4E-02	22	0.96	3.22
16	3,5,7	8	8.5E-04	5	0.14	5.93	2,3,7,8	5	2.4E-03	5	0.63	3.70	3,5	2	4.0E-02	62	0.93	3.12
17	5,8	7	1.5E-03	9	0.69	5.87	5,7	3	1.7E-02	36	0.43	3.65	2,8	1	9.0E-03	14	0.61	3.11
18	**14**	**12**	**6.1E-03**	**36**	**0.24**	**5.84**	1,7,8	3	2.9E-03	6	0.43	3.61	1,2	3	3.5E-02	54	0.30	3.06
19	**12**	**14**	**6.1E-03**	**36**	**0.15**	**5.84**	2,5,7	3	1.6E-02	33	0.42	3.56	1,2,7	3	1.9E-02	29	0.29	3.03
20	1,5,7	8	1.0E-03	6	0.14	5.82	1,2,5	3	2.4E-02	50	0.42	3.51	10,3	2	5.8E-03	9	0.90	3.03

	$V_c \cap P_{>1d}$ (531 devices)						$V_c \cap P_{>1d} \cap H$ (196 devices)					
	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l
1	1,2,7,8	3	1.1E-02	6	0.75	4.06	1,7,8	3	2.6E-02	5	0.83	3.02
2	2,5,8	3	9.4E-03	5	0.71	3.87	1,5,7	3	2.6E-02	5	0.83	3.02
3	2,7,8	3	1.3E-02	7	0.70	3.79	1,2,7,8	3	2.6E-02	5	0.83	3.02
4	1,2,5,7	3	2.6E-02	14	0.70	3.79	1,2,5,7	3	2.6E-02	5	0.83	3.02
5	2,5,7	3	3.4E-02	18	0.69	3.75	5,7	3	4.1E-02	8	0.80	2.90
6	1,5,7	3	2.6E-02	14	0.67	3.61	2,5,7	3	4.1E-02	8	0.80	2.90
7	5,7	3	3.6E-02	19	0.66	3.55	7,8	3	3.1E-02	6	0.75	2.72
8	1,7,8	3	1.1E-02	6	0.60	3.25	2,7,8	3	3.1E-02	6	0.75	2.72
9	5,8	3	9.4E-03	5	0.56	3.01	12,2	3	3.1E-02	6	0.60	2.18
10	1,2,5	3	4.9E-02	26	0.53	2.88	2,5	3	1.1E-01	21	0.60	2.18
11	2,5	3	6.8E-02	36	0.53	2.87	1,3,7	8	2.6E-02	5	0.21	2.15
12	3,8	5	9.4E-03	5	0.45	2.51	1,2,3,7	8	2.6E-02	5	0.21	2.15
13	2,3,8	5	9.4E-03	5	0.45	2.51	2,3,7	8	3.1E-02	6	0.21	2.13
14	1,2,7	3	7.5E-02	40	0.45	2.46	2,3	5	1.1E-01	21	0.42	2.06
15	1,2,8	3	1.7E-02	9	0.45	2.44	1,2,3	8	3.6E-02	7	0.19	2.01
16	1,5	3	4.9E-02	26	0.45	2.43	3,7	8	3.1E-02	6	0.19	2.00
17	2,3	5	6.8E-02	36	0.42	2.34	1,3	8	3.6E-02	7	0.19	1.95
18	1,2,3	5	4.9E-02	26	0.42	2.32	**5**	**3**	**1.1E-01**	**21**	**0.53**	**1.91**
19	2,8	3	2.1E-02	11	0.42	2.29	**3**	**5**	**1.1E-01**	**21**	**0.39**	**1.91**
20	5,7	8	1.1E-02	6	0.21	2.29	1,2,5	3	6.1E-02	12	0.52	1.89

Table 5.4: Top-20 (where applicable) of association rules $X \Rightarrow Y$ for visitor segments $V \cap H$, $V \cap H_{far}$, $V \cap H_{4*}$, and $V \cap H_{hostel}$. Rules were generated based on constraints on both support ($s_a \geq 5$) and confidence ($c \geq 0.05$), and are sorted on their lift (l) value. The relative (s_r) and absolute (s_a) support, as well as the confidence (c) are also shown. Rules with only item in the antecedent (marked in bold) are also visible in figure 5.8 on page 108.

	$V \cap H$ (483 devices)						$V \cap H_{far}$ (71 devices)						$V \cap H_{4*}$ (249 devices)					
	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l
1	1,3,7	8	1.00E-02	5	0.21	5.3	1,2	7	7.00E-02	5	0.21	1.48	1,3,7	8	2.00E-02	5	0.36	7.41
2	1,2,3,7	8	1.00E-02	5	0.21	5.3	**3**	**2**	**2.10E-01**	**15**	**0.88**	**1.36**	1,2,3,7	8	2.00E-02	5	0.36	7.41
3	2,3,7	8	1.20E-02	6	0.21	5.26	**2**	**3**	**2.10E-01**	**15**	**0.33**	**1.36**	2,3,7	8	2.40E-02	6	0.33	6.92
4	3,7	8	1.20E-02	6	0.19	4.92	1,3	2	9.90E-02	7	0.88	1.35	3,7	8	2.40E-02	6	0.32	6.55
5	2,5	8	1.00E-02	5	0.14	3.53	**7**	**2**	**1.10E-01**	**8**	**0.8**	**1.23**	1,2,7	8	2.00E-02	5	0.23	4.72
6	1,7,8	3	1.00E-02	5	0.83	3.53	**2**	**7**	**1.10E-01**	**8**	**0.17**	**1.23**	7,8	3	2.40E-02	6	1	4.61
7	1,5,7	3	1.00E-02	5	0.83	3.53	1,2	3	9.90E-02	7	0.29	1.22	1,7,8	3	2.00E-02	5	1	4.61
8	1,2,7,8	3	1.00E-02	5	0.83	3.53	1,7	2	7.00E-02	5	0.71	1.1	2,7,8	3	2.40E-02	6	1	4.61
9	1,2,5,7	3	1.00E-02	5	0.83	3.53	**7**	**1**	**9.90E-02**	**7**	**0.7**	**1.08**	1,2,7,8	3	2.00E-02	5	1	4.61
10	5,7	3	1.70E-02	8	0.8	3.39	**1**	**7**	**9.90E-02**	**7**	**0.15**	**1.08**	2,5	8	2.00E-02	5	0.22	4.51
11	2,5,7	3	1.70E-02	8	0.8	3.39	2,7	1	7.00E-02	5	0.63	0.96	1,2,3	8	2.80E-02	7	0.21	4.4
12	1,3,8	7	1.00E-02	5	0.71	3.35	**1**	**2**	**3.40E-01**	**24**	**0.52**	**0.81**	2,8	5	2.00E-02	5	0.45	4.19
13	1,2,3,8	7	1.00E-02	5	0.71	3.35	**2**	**1**	**3.40E-01**	**24**	**0.52**	**0.81**	1,3	8	2.80E-02	7	0.2	4.15
14	1,2,7	8	1.20E-02	6	0.13	3.25	**3**	**1**	**1.10E-01**	**8**	**0.47**	**0.73**	5,7	3	2.40E-02	6	0.86	3.95
15	2,8	5	1.00E-02	5	0.28	3.19	**1**	**3**	**1.10E-01**	**8**	**0.17**	**0.73**	2,5,7	3	2.40E-02	6	0.86	3.95
16	7,8	3	1.20E-02	6	0.75	3.18	2,3	1	9.90E-02	7	0.47	0.72	**8**	**5**	**2.00E-02**	**5**	**0.42**	**3.84**
17	2,7,8	3	1.20E-02	6	0.75	3.18							**5**	**8**	**2.00E-02**	**5**	**0.19**	**3.84**
18	2,3,7	5	1.70E-02	8	0.28	3.17							2,3	8	3.20E-02	8	0.17	3.61
19	3,8	7	1.20E-02	6	0.67	3.13							2,7	8	2.40E-02	6	0.17	3.56
20	2,3,8	7	1.20E-02	6	0.67	3.13							3,8	7	2.40E-02	6	0.75	3.4

	$V \cap H_{hostel}$ (25 devices)																	
	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l	X	Y	s_r	s_a	c	l
1	1,2	5	1.20E-01	3	0.43	3.57												
2	**5**	**2**	**1.20E-01**	**3**	**1**	**2.08**												
3	**2**	**5**	**1.20E-01**	**3**	**0.25**	**2.08**												
4	1,5	2	1.20E-01	3	1	2.08												
5	**5**	**1**	**1.20E-01**	**3**	**1**	**1.56**												
6	**1**	**5**	**1.20E-01**	**3**	**0.19**	**1.56**												
7	2,5	1	1.20E-01	3	1	1.56												
8	**2**	**1**	**2.80E-01**	**7**	**0.58**	**0.91**												
9	**1**	**2**	**2.80E-01**	**7**	**0.44**	**0.91**												

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, volume 22, pages 207–216, New York, USA. ACM Press.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499, Santiago.

Ahas, R., Aasa, A., Mark, U., Pae, T., and Kull, A. (2007a). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898–910.

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Ahas, R., Laineste, J., Aasa, A., and Mark, U. (2007b). The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In Gartner, G., Cartwright, W., and Peterson, M. P., editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 445–460. Springer, Berlin.

Al-Salim, B. (2008). Mass customization of travel packages: data mining approach. *International Journal of Flexible Manufacturing Systems*, 19(4):612–624.

Appice, A. and Buono, P. (2005). Analyzing Multi-level Spatial Association Rules Through a Graph-Based Visualization. In Moonis, A. and Esposito, F., editors, *Innovations in Applied Artificial Intelligence*, volume 3533 of *Lecture Notes in Computer Science*, pages 448–458. Springer.

Bloom, J. Z. (2005). Market Segmentation: A Neural Network Application. *Annals of Tourism Research*, 32(1):93–111.

Bonné, B., Barzan, A., Quax, P., and Lamotte, W. (2013). WiFiPi: Involuntary tracking of visitors at mass events. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Madrid.

Bruzzese, D. and Davino, C. (2003). Visual post-analysis of association rules. *Journal of Visual Languages & Computing*, 14(6):621–635.

Burger, C., Dohnal, M., Kathrada, M., and Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting — a case study of Durban, South Africa. *Tourism Management*, 22(4):403–409.

Chen, Y.-L., Tang, K., Shen, R.-J., and Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2):339–354.

Cini, F., Leone, L., and Passafaro, P. (2010). Promoting Ecotourism Among Young People: A Segmentation Strategy. *Environment and Behavior*, 44(1):87–106.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., and Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages & Computing*, 18(3):255–279.

Connell, J. and Page, S. J. (2008). Exploring the spatial patterns of car-based tourist travel in Loch Lomond and Trossachs National Park, Scotland. *Tourism Management*, 29(3):561–580.

Delafontaine, M., Versichele, M., Neutens, T., and Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34:659–668.

Dolničar, S. (2004). Beyond "Commonsense Segmentation": A Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*, 42(3):244–250.

Dolničar, S. and Leisch, F. (2003). Winter tourist segments in Austria - Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research*, 41(3):281–292.

Eagle, N. and Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Emel, G. G., Taskin, c., and Akat, O. (2007). Profiling a Domestic Tourism Market By Means Of Association Rule Mining : A Case Study. *Anatolia*, 18(2):1–12.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI magazine*, 17(3):37–54.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hahsler, M. and Chelluboina, S. (2010). Visualizing Association Rules : Introduction to the R-extension Package arulesViz. *R project module*.

Hahsler, M., Chelluboina, S., Hornik, K., and Buchta, C. (2011). The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets. *Journal of Machine Learning*, 12:2021–2025.

Hartmann, R. (1988). Combining field methods in tourism research. *Annals of Tourism Research*, 15:88–105.

Hong, W.-C., Dong, Y., Chen, L.-Y., and Wei, S.-Y. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing*, 11(2):1881–1890.

Janelle, D., Goodchild, M., and Klinkenberg, B. (1988). Space-time diaries and travel characteristics for different levels of respondent aggregation. *Environment and Planning A*, 20:891–906.

Kemperman, A., Borgers, A., and Timmermans, H. (2009). Tourist shopping behavior in a historic downtown area. *Tourism Management*, 30(2):208–218.

Lau, G. and McKercher, B. (2006). Understanding tourist movement patterns in a destination: A GIS approach. *Tourism and Hospitality Research*, 7(1):39–49.

Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management*, 21(4):331–340.

Law, R. and Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management*, 20:89–97.

Law, R., Bauer, T., Weber, K., and Tse, T. (2006). Towards a Rough Classification of Business Travelers. In *Advanced Data Mining and Applications*, volume 4093 of *Lecture Notes in Computer Science*, pages 135–142.

Law, R., Rong, J., Vu, H. Q., Li, G., and Lee, H. A. (2011). Identifying changes and trends in Hong Kong outbound tourism. *Tourism Management*, 32(5):1106–1114.

Lew, A. and McKercher, B. (2006). Modeling Tourist Movements. *Annals of Tourism Research*, 33(2):403–423.

Liao, S.-h., Chen, Y.-J., and Deng, M.-y. (2010). Mining customer knowledge for tourism new product development and customer relationship management. *Expert Systems with Applications*, 37(6):4212–4223.

Lonely Planet (2011). Ghent: Belgium's best kept secret.

McKercher, B. (1999). A chaos approach to tourism. *Tourism Management*, 20:425–434.

Min, H., Min, H., and Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. *International Journal of Contemporary Hospitality Management*, 14(6):274–285.

Orellana, D., Bregt, A. K., Ligtenberg, A., and Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management*, 33(3):672–682.

Öztayşi, B., Baysan, S., and Akpinar, F. (2009). Radio frequency identification (RFID) in hospitality. *Technovation*, 29(9):618–624.

Palmer, A., José Montaño, J., and Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management*, 27(5):781–790.

Peterson, B., Baldwin, R., and Kharoufeh, J. (2006). Bluetooth inquiry time characterization and selection. *IEEE Transactions on Mobile Computing*, 5(9):1173–1187.

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

Rong, J., Vu, H. Q., Law, R., and Li, G. (2012). A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining. *Tourism Management*, 33(4):731–740.

Shoval, N. and Isaacson, M. (2007a). Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*, 97(2):282–297.

Shoval, N. and Isaacson, M. (2007b). Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1):141–159.

Shoval, N. and Isaacson, M. (2009). *Tourist mobility and advanced tracking technologies*, volume 19 of *Routledge Advances in Tourism*. Routledge, New York, London.

Shoval, N., McKercher, B., Ng, E., and Birenboim, A. (2011). Hotel location and tourist activity in cities. *Annals of Tourism Research*, 38(4):1594–1612.

Song, H. and Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism Management*, 29(2):203–220.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Association Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining*, page 769. Addison-Wesley Longman Publishing, Boston, MA, USA.

Tchetchik, A., Fleischer, A., and Shoval, N. (2009). Segmentation of Visitors to a Heritage Site Using High-resolution Time-space Data. *Journal of Travel Research*, 48(2):216–229.

Techapichetvanich, K. and Datta, A. (2005). VisAR : A New Technique for Visualizing Mined Association Rules. In Xue, L., Wang, S., and Zhao Yang, D., editors, *Advanced Data Mining and Applications*, volume 3584 of *LNCS*, pages 88–95.

Toerisme Vlaanderen (2012). Toerisme in cijfers 2011: de Belgische markt in Vlaanderen. Technical report, Toerisme Vlaanderen.

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012a). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2):208–220.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., and Van de Weghe, N. (2012b). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10):14196–14213.

Witt, S. F. and Witt, C. a. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11(3):447–475.

World Tourism Organization (1995). Collection of Tourism Expenditure Statistics. Technical report, World Tourism Organization.

Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):60–70.

6 Discussion and conclusions

Crowds and the movements of individuals constituting them hold an important relevance in a variety of domains. Empirical movement data are generally considered to be of paramount importance, both for direct analyses as well as the validation of simulation models. This dissertation set off by focusing on the challenging nature of gathering such empirical data, and how new developments such as the emergence of positioning technologies like GPS offer new opportunities in comparison with conventional methodologies. Mobile phones are increasingly considered as the main catalyst of this process. As they are in close contact with their owners, they can be used as 'proxies' for measuring human movement. Generally, there are two established approaches for analyzing movement patterns through mobile phones. The first involves users actively sharing their (GPS-based) location through smartphone applications (e.g. Facebook, Foursquare, Twitter, Flickr). A second approach is to reconstruct phone movements from call records of mobile operators. This non-participatory method (users are not involved nor aware of the experiment) seems better suited for studying large groups of participants, but the spatiotemporal level of detail of the resulting trajectories is usually insufficient for smaller-scale movements.

Recently, it has been proposed that short-range wireless technologies, such as WiFi or RFID but Bluetooth in particular, could be used in a manner which also allows to trace movements of mobile phones (for example at mass events). More specifically, trajectories can be reconstructed by mapping detections of the same mobile device to different strategically placed Bluetooth sensors through the device's MAC address (which acts a unique identifier of the device). This dissertation sought to build upon this new approach and its limited body of academic literature to date. In doing so, the aim was to [i] further illustrate and document the benefits and issues of Bluetooth tracking at mass events; [ii] explore the potential for applications outside the scope of mass events; and [iii] investigate the process of analyzing Bluetooth tracking data and their specific characteristics. An application-oriented approach was adopted, where all research questions were empirically addressed in different case studies.

This concluding chapter starts by summarizing the main achievements reported in this dissertation (section 6.1). This summary is followed by a discussion (section 6.2 on page 126) reflecting on the main contributions of the presented research (section 6.2.1 on page 126) and the remaining issues that deserve further attention in the future (section 6.2.2 on page 129). More specifically, we discuss three important issues of Bluetooth tracking: the lack of attributes on anonymous tracked individuals (section 6.2.2.1 on page 129), the representativity of the tracked population (section 6.2.2.2 on page 130) and privacy issues (section 6.2.2.3 on page 134). As a reference for discussing representativity, we will also shortly reflect on our first experiences with 'WiFi tracking' — which is a similar methodology as Bluetooth tracking but based on detecting WiFi devices instead of Bluetooth devices. We end this chapter with the main conclusions of the dissertation (section 6.3 on page 136).

6.1 Summary

Based on the scientific literature and the main objectives of the dissertation outlined in the introduction, we distilled a research agenda composed of four research questions. This section summarizes the main achievements of chapters 2–5 in addressing these research questions, and of the *GISMO* toolkit presented in appendix A on page 141.

RQ 1: Which opportunities does Bluetooth tracking provide for studying spatiotemporal dynamics within crowds at mass events?

Due to its non-participatory nature, Bluetooth tracking seems particularly promising for capturing the mobility of large crowds at mass events. Despite a limited number of documented use-cases (Larsen et al., 2013; Leitinger et al., 2010; Stange et al., 2011), the methodology remains relatively little-known and lacks a foundation of academic literature. Chapter 2 on page 19 addressed the first research question by illustrating the potential of Bluetooth tracking at the 'Ghent Festivities 2010' event, which attracted around 1.5 million visitors over ten days. In contrast with existing literature, the chapter aimed to give a broad overview of all analytical possibilities on Bluetooth tracking data instead of focusing on one specific problem. We started by presenting the Bluetooth tracking methodology in detail, focusing both on the working principle as well as on the hardware (sensors) deployed during the experiment. By installing Bluetooth sensors at 22 locations (11 of which inside the event zone; the others at points of entry, two train stations and a park&ride tram stop) a large dataset was gathered which contained 152,487 trajectories of 80,828 mobile devices (phones) detected within the confines of the event. The information potential residing within the dataset was subsequently illustrated by performing different analyses.

First, we calculated that 11.0±1.8% of the public was detected through a Bluetooth device by comparing detected devices with visual head counts at different locations. Through this 'detection ratio', we roughly estimated the total number of visitors at 1.4 million (minimum: 1.2 million, maximum: 1.7 million). This was in line with the expectations of the event organizers. By aggregating the data of different sensors located inside the event zone, we were able to reconstruct the varying amount of visitors over time. Details such as crowded and less crowded days, and different profiles of crowdedness over time for the different days were identified and interpreted. Additionally, a pattern was discovered where the crowd is generally spread out over most of the center during the day but concentrates around one square after midnight. Further analysis indicated that the majority (65%) of visitors were one day visitors. The shares of several day visitors over the different squares separately varied significantly between 8% and 20%. We also detected visitors at both train stations and the largest park&ride tram stop of the city. We deduced that the share of visitors that took the train was more or less constant over the different days (5–6%), but that the share of tram users varied more (3–7%). By detecting visitors at points of entry, we calculated visit durations. The median value was around 3.5 hours, but the distribution showed a large spread with a heavy tail towards longer visits (11% of the sample stayed for at least 7 hours). A concise flow analysis revealed a typical pattern of crowd concentration after midnight, followed by a general efflux. The chapter was concluded with a discussion reflecting on the merits and remaining issues of the Bluetooth tracking methodology. The main identified added value in comparison with conventional methodologies was its ubiquitous applicability (indoor + outdoor), its relatively straightforward way of generating large mobility datasets, and its non-participatory nature. However, important questions related to the representativity of the sampled set of tracked individuals remained.

RQ 2: Can Bluetooth technology be used to count and map complex crowds dispersed over large areas?

Where a crowd and its dynamics within the well-defined confines of an event can be measured through a distributed network of Bluetooth sensors, the approach of using static sensors ultimately fails in studying crowds spread out over a large geographic extent. Chapter 3 on page 45 addressed this research question by studying the crowd of spectators at the 'Tour of Flanders 2011' cycling race, which was spread out over a 257 kilometer long race track. We employed a mobile mapping approach, where a mobile platform (car) equipped with two Bluetooth sensors and preceding the racers between 3 and 6 minutes mapped the spectators as it passed them by. A preliminary experiment indicated that the speed of the platform negatively influenced the detection process, but that class 1 Bluetooth sensors (the most sensitive class) did not miss any of the mobile devices on the side of the road on any run. During the actual experiment, around 16,000 mobile phones were detected and mapped to positions along the track. By dividing the trajectory into 1 kilometer long segments, we generated a detailed map of the relative crowdedness along the race track where hotspots usually corresponded to slopes or cobblestoned segments. By visually counting spectators (through a camera on the mobile platform) and comparing these counts with numbers of detected Bluetooth devices, we estimated a detection ratio of 14.3$\pm$3.9% with outliers and 13.0$\pm$2.3% without outliers. Through extrapolation based on these figures, we estimated the amount of spectators over the entire trajectory and over the crowdiest segment. A subsequent analysis did not provide indications of a direct influence of the platform speed on a number of indicators of the detection process, but the relatively low overlap (20–80%) between both sensors on the mobile platform demonstrated the importance of the exact placement of sensors. The relative standard error of 17.9% on the detection ratio was slightly higher than that cited for estimations from photographic images $\pm$10% according to Watson and Yip (2011). The main advantage of the demonstrated proof-of-concept over conventional methods, however, does not solely lie in the accuracy of counting a crowd at a fixed moment in time, but in the ability of automatically providing a view on the spatiotemporal evolution of crowds based on individual trajectories.

RQ 3: How can a crowd's movement within sensor locations be modeled?

This research questions stems from the spatiotemporally sparse or 'episodic' (Andrienko et al., 2012) nature of Bluetooth tracking data. As devices are only detected at locations where a sensor is present, trajectories are usually characterized by considerable periods of time where the location is unknown due to the device being out of range of all sensors. The potential path between these two locations, might, however be modeled. Chapter 4 on page 67 addresses this research question by linking with concepts from the field of time geography. Based on the space-time prism construct, a model for calculating the potential co-presence of several moving objects over a network is developed. By limiting the deviation from the shortest path between two sensor locations, the model constrains to feasible co-presence opportunities. The model was applied to a Bluetooth tracking dataset containing trajectories of visitors of the 'Ghent Light Festival 2012', which showcased 29 artworks along a 'light trajectory' for four days. By visualizing a time-series of the model output, we were able to reconstruct the crowd flow over the entire study area despite only 25 Bluetooth sensors being deployed. Subsequently, the value of these maps of feasible co-presences was discussed by reflecting on the difference between actual co-presence and potential co-presence. As a concise first validation of the model, the relationship between these two variables was examined for two sensors used within the model and one sensor which was not included in the model. The output of the model could be used as an indicator of potential crowdedness, but further validation was identified as a key aspect of future work.

RQ 4: What is the value of Bluetooth tracking outside of the context of mass events?

Where the previous research questions investigated crowds as spatiotemporally aggregated groups, this research question broadened the focus to crowds as groups of individuals with similar intentions. Chapter 5 on page 89 addressed the research question by translating it to a tourism context. Over the course of two weeks in 2012, visitors to 14 of the most important tourist attractions in Ghent were registered by Bluetooth sensors. Additionally, some visitors were also identified as hotel guests by sensors deployed in 14 hotels in and around the city and as inquirers by a sensor installed in the tourist inquiry desk. We first demonstrated that the deduction of activities (making a visit, staying at a hotel) from tracking data requires an extensive filtering procedure in order to remove data noise. Subsequently, we first identified several visitor segments based on *inter alia* the difference between visits to open and closed attractions (the difference being in the need for registration upon entering) and the identification as a hotel guest. An association rule learning scheme was applied to each of these different visitor segments in order to discover interesting associations between different attractions. The generated information was visualized by means of 'visit pattern maps', which combined a geographical depiction of the discovered association rules and the spatial distribution of visits over the different attractions. Despite the need for filtering and the limited tracking period, the combination of Bluetooth tracking and a data mining technique such as association rule learning was able to generate valuable insights. This knowledge could be used in the short term for strengthening existing or creating new associations between different attractions by directed advertising, or in recommendation systems based on collaborative filtering. In the longer term, visit pattern maps could help urban planners in improving tourist facilities. The concluding discussion further reflected on the need for filtering, the unknown representativity of the data, and the lack of other attributes of tracked individuals other than their location over time.

Due to the novelty of Bluetooth tracking, analysis of the resulting data is hampered by a lack of specialized software. Existing tools and methods are ill-suited for analyzing Bluetooth tracking data due to their episodic nature (see also research question 3). Appendix A on page 141 addressed this issue by presenting the *GISMO* toolkit, which was developed over the last four years in order to process Bluetooth tracking data but can also handle other types of episodic movement data. The different capabilities of the toolkit were illustrated by preprocessing, analyzing and visualizing a tracking dataset gathered at a music festival. These capabilities included making selections (based on device type, brand, spatial or spatiotemporal constraints), live data filtering, sampling of a variety of trajectory/device properties, calculation of flows, and visualization through graphs or by exporting (time-aware) KML files for further exploration in Google Earth. The toolkit's main merit lies in assembling a number of common procedures in preprocessing, selecting, transforming, analyzing and visualizing episodic proximity-based movement data — and in particular Bluetooth tracking data — under one accessible user interface. Additionally, the toolkit served an important supportive role for this dissertation by being used for analyses in chapters 2, 4 and 5.

6.2 Discussion

The aim of this section is to critically reflect upon the results presented throughout this dissertation and summarized in the previous section. Doing so, it serves as a compilation of the most important points addressed in the discussion sections of the separate chapters, supplemented by additional global insights after four years of research and recent preliminary findings which have not yet been published. We start by outlining the main scientific contributions of the dissertation, and follow up by discussing the most important remaining issues regarding the use of Bluetooth tracking data.

6.2.1 Main contributions

As discussed in the introduction, the application of Bluetooth technology for gathering mobility data represents a very young field of research. After the first reported commercial application in 2002 (by the Danish company Bluelon[1], for calculating queuing times in airports), it took a few years before academic interest started to surface. Where the first initiatives mainly focused on participatory interaction modeling (Eagle and Pentland, 2005; Hui et al., 2005), recent years have demonstrated a growing interest in using the technology for the non-participatory tracking of large crowds at mass events (Larsen et al., 2013; Leitinger et al., 2010; Stange et al., 2011). In parallel with these scientific investigations, the methodology is gaining particular attention for its economical perspectives. A growing number of start-up companies use the approach to generate valuable knowledge from mobility patterns in retail environments[2]. This rapid expansion of applications represents a characteristic example of the way ICT, and more specifically wireless and location-aware technologies, are predicted to fundamentally change our world. Without going into details, it is clear that Bluetooth tracking cannot be seen separately from the more general trend of new paradigms in ICT-related research such as Ubiquitous Computing (Abowd and Mynatt, 2000), the Internet of Things (Atzori et al., 2010) and in particular Smart Cities (Schaffers et al., 2011).

In comparison to the rapid growth in commercial Bluetooth tracking solutions, the body of scientific literature on the methodology and its use (in particular for pedestrian mobility) remains rather limited. Most reported use-cases describe concise proof-of-concept demonstrations (Leitinger et al., 2010; Stange et al., 2011), or use Bluetooth tracking data for performing specific analyses (Larsen et al., 2013) or illustrating specific methods (Andrienko et al., 2012). In contrast, chapter 3 on page 45 aimed to focus on the methodology itself and give a comprehensive overview of its added value in the context of mass events. The

[1] www.bluelon.com
[2] BlipSystems, Renew-PresenceOrb, ShopperTrak, etc.

main contribution of the case study at the 'Ghent Festivities 2010' — which, to the best of our knowledge, represented the largest Bluetooth tracking experiment to that date — is that it provided a complete image of the information potential by empirical analyses on the gathered dataset. The approach was to describe the entire architecture (both hard- and software) as well as the management and processing of the tracking data in such detail that the methodology became more tangible to a general audience. This was a particularly relevant aim as only few event organizers, in our own experience, are aware of the potential or even the existence of Bluetooth tracking.

Where the contribution of this dissertation in the case of static sensors was to further scientifically establish an innovative yet existing practice, the mobile mapping of a dispersed crowd with Bluetooth sensors presented in chapter 3 on page 45 is completely novel. To the best of our knowledge, there have been no previous mobile mapping applications making use of Bluetooth technology for studying crowds. The use of smartphones as wearable Bluetooth sensors *has* very recently been documented at the Roskilde festival (Stopczynski et al., 2013), but our mobile mapping approach differs in two important ways. First, it does not require the active participation of festival visitors, which avoids the risk of heterogeneous data quality. Second, it uses high-quality Bluetooth sensors with external antennas in comparison to the less sensitive sensors built into current smartphones. As a result, these sensors can track individuals which are not in the immediate vicinity of the mobile platform trajectory. As the mobile platform stayed relatively close to the cyclists, it can be expected that the majority of spectators were detected. Nevertheless, people lining up next to the race track after the mobile platform had passed (giving them between 3 to 6 minutes to see the first cyclists) were not detected. Other scenarios of crowds lacking a well-defined attractor (in this case the cyclists) will necessitate additional mobile sensors deployed over the study area. In case some zones would not be directly accessible to land-based platforms, airborne platforms such as unmanned aerial vehicles (UAV) might be employed.

In chapter 4 on page 67, a model was developed for calculating the feasible co-presence opportunities of agents moving over a network and being detected at certain locations on this network. The concept was subsequently applied to a Bluetooth tracking dataset gathered at the 'Ghent Light Festival', and used to generate time-series of maps depicting the modeled dispersion of the crowd in between sensor locations. By linking to the field of time geography, the model built on the growing body of research concerned with joint accessibility (Miller, 2005; Neutens et al., 2008, 2007) and social interaction potential (Farber et al., 2013; Neutens et al., 2013), and the increasing availability of toolkits and implementations for calculating interaction spaces (Fang et al., 2011; Kang and Scott, 2007; Neutens et al., 2010). Rather than introducing new theoretical concepts, the main novelty and contribution of the chapter is in the use of interaction potential as an indicator of potential crowdedness and in the application of time-geographical concepts to Bluetooth tracking data. The model's utility is not confined to Bluetooth tracking data as such, as it can handle any type of episodic movement data (Andrienko et al., 2012). As these datasets will become more widespread due to the emergence of new approaches in mobility measurement, we envision a corresponding growing interest in modeling movements and/or interactions in between sparse location registrations. Mobile phone datasets (Blondel et al., 2010; Candia et al., 2008; González et al., 2008) represent one such example where the model could open up a wide array of new analytical possibilities. While it can be expected that the interpretability of the model output will be less straightforward in scenarios with less coordination between individual movements than in the case of the 'Ghent Light Festival', the application of the model in a broader scope of scenarios is left for future work.

By shifting the focus to a tourism context in chapter 5 on page 89, we aimed to demonstrate that Bluetooth tracking has applications besides those situated at mass events. The chapter describes the very first use of Bluetooth technology for tracking tourist movements, illustrating the value of Bluetooth tracking as a methodological alternative to conventional approaches such as the use of GPS technology (Shoval and Isaacson, 2009). We demonstrated that the methodology holds particular promise for studying sub-regional movements of large samples of individuals. An important distinction from the investigations in the previous chapters is that the basic unit of analysis was changed from presences to activities (in this case performing a visit or staying at a hotel). Inferring of activities from Bluetooth proximity detections at strategic locations holds great potential in further extending the range of analytical possibilities. To the best of our knowledge, however, only one study has reported such an approach. In the Roskilde festival, visitors were inferred to have attended a performance when being detected a certain minimum number of times by a sensor near the stage a band was playing on (Larsen et al., 2013). As described in chapter 5 on page 89, identifying attraction visitors or hotel guests in a busy city is significantly more challenging. Due to the sensors being deployed in public spaces, there was a large amount of detected individuals merely passing by sensor locations. The detailed description of the filtering process needed to remove this data noise represents another important contribution of the chapter, although different approaches might be used in the future. An association rule learning scheme was applied to the filtered dataset, and the generated rules were geographically plotted on 'visit pattern maps'. These visit pattern maps represent a third significant contribution in that they summarize information which would otherwise be harder to interpret in the usual tabular manner, and could therefore function as important documents for policy makers.

Besides the important role of facilitating analyses performed in Chapters 2, 4 and 5, the main contribution of the *GISMO* toolkit presented in appendix A on page 141 is that it represents the first toolkit specifically tailored for analyzing (Bluetooth-based) episodic movement data. Due to the growing importance of these movement data, toolkits such as *GISMO* will become invaluable tools to academics and policy-makers. The user-friendly interface might impose some limitations in comparison with more general analytical or data mining solutions, but opens up Bluetooth tracking to a wider audience which might not be accustomed to working with software characterized by steeper learning curves. The use of the toolkit in ten Masters' theses besides this dissertation (see table B.1 on page 167) can serve as a case in point.

6.2.2 Remaining issues

Over the course of the different studies reported in this dissertation, several issues of Bluetooth tracking and Bluetooth tracking data were identified which need further investigation. This section reflects on three of these issues which bear relevance over the entire scope of the dissertation: the lack of attributes in Bluetooth tracking data, the representativeness of Bluetooth tracking data with reference to the entire studied population, and privacy implications.

6.2.2.1 Lack of attributes in anonymous tracking data

As described in chapter 2 on page 19, Bluetooth tracking data are anonymous in the sense that the only information on tracked individuals besides their location over time is the unique hardware identifier (MAC address) of the device they are carrying. The lack of any other personal attributes such as gender, age, social background, etc. prohibits the testing of hypotheses on why certain movements or activities are performed. The importance of explaining variables becomes particularly relevant outside of the context of mass events. Many studies on tourist spatiotemporal behavior, for example, make intensive use of demographic or psychographic variables. Due to the lack of such variables, studies based on Bluetooth tracking data alone are limited to observing certain behavior without the ability of fully explaining it. Some of the patterns discovered in the case study of chapter 5 on page 89, for example, could be caused by tourists with specific characteristics. Although some assumptions could be made (e.g. tourists staying in hostels are younger than tourists staying in four-star hotels), they would be hard to validate without further data on tracked individuals. The need for making assumptions in certain analyses, in general, represents an additional concern when working with anonymous data — further impeded by the lack of an empirical foundation for validation of these assumptions. As a consequence, the assumptions made in this dissertation were often based on a combination of intuition, experience (when working with the gathered data) and common sense.

So, the non-participatory approach in measuring the spatiotemporal behavior of individuals offers clear advantages with regard to scalability but is limited in explaining individual behavior. This is in sharp contrast with conventional methods directly approaching study objects, which are usually rich in individual details but poor in coverage of the studied population. As such, this contrast links to the traditional debate on the (in)compatibility of qualitative and quantitative methods (Howe, 1988) and on the growing recognition that both types of approaches should be combined (Fielding and Schreier, 2001). This recognition is also crystallizing in the field of pedestrian behavior monitoring (Millonig and Gartner, 2007). Bluetooth tracking data could, for example, be combined with personal interviews. How such a combination should be achieved is not immediately clear, however. First, interviews and the tracking data would be linked by registering the MAC address of interviewees. As only a fraction of these interviewees will be detectable, a large number of interviews will need to be performed in order to contact an acceptable number of tracked individuals. This entails the risk of nullifying the main advantage of non-participatory tracking, namely its less labor-intensive nature and its ability to scale well with crowd size. An appropriate balance will need to be found. Second, contacted individuals are usually not familiar with the concept of Bluetooth tracking and might have reservations cooperating due to the potential privacy implications of linking personal data to their trajectories (see section 6.2.2.3 on page 134). Third, there is a need for a quick and practical way of making an exact match between an interviewee and the MAC address of his/her device. One approach could be to check for the friendly name of the device, but these are by default set to general descriptors of the device (usually the brand followed by the device type). Making the match solely through proximity is problematic if the interview is taken in a public space containing other detectable mobile devices. In sum, future experiments should further investigate how to combine Bluetooth tracking with other qualitative methods.

6.2.2.2 Representativeness of Bluetooth tracking data

Besides impeding certain analyses, the lack of additional information on tracked individuals makes it impossible to verify how representative Bluetooth tracking data are with regards to the entire studied population. In this section, we critically reflect on two different yet related aspects of this representativeness: the detection ratio and the danger of self-selection bias.

Detection ratio The detection ratio represents the share of individuals in a crowd which are detected through a mobile device. As it gives a first rough idea of the representativeness of the gathered tracking data (a higher detection ratio will lower the risk of studying unrepresentative subsets of individuals), this value has been calculated for several Bluetooth tracking projects over the last five years. A number of counting methods have been used over the years: manual counting (on the field), using optical barriers, and using video imagery (both manually and automatically). The overview given in figure 6.1 on the next page hints at an important trend where it seems that the Bluetooth detection ratio has progressively decreased from around 10% in 2009–2010 to between 5 and 10% in 2013. Besides the detection ratios, we have also shown the relative standard errors (RSE) associated with the detection ratios. These represent the relative amount of variance around the calculated mean, and as such reflect the error made when extrapolating numbers of detected devices to actual crowd counts. At first sight, it might seem counter-intuitive that detection ratios calculated from larger samples lead to larger RSE values. However, relative standard errors of samples with similar standard deviations increase with decreasing mean values. Additionally, the two detection ratios with the largest sample sizes were based on automatic people counting systems which functioned for ten days and provided detection ratio estimations on an hourly basis. As a result, they might better capture short-term temporal variations in detection ratios.

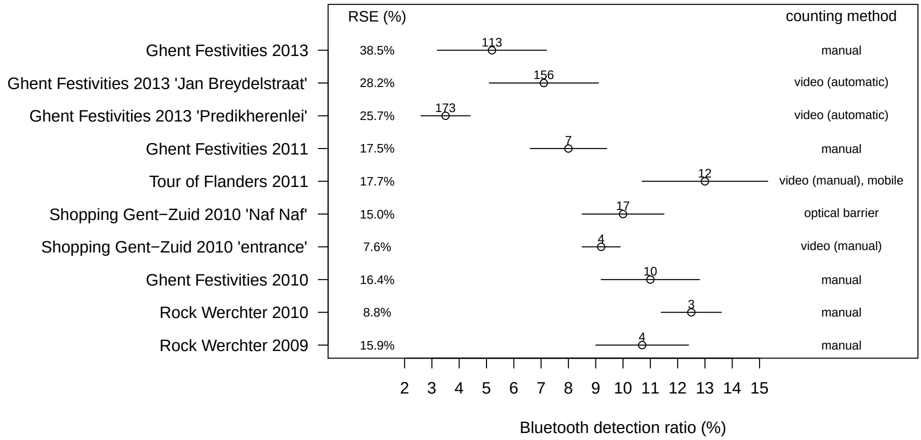


Figure 6.1: Overview of Bluetooth detection ratios calculated over the last five years. Circles represent the mean values, horizontal bars the standard deviations, and labels above the circles the sample sizes (i.e. the number of periods over which counts were made). The relative standard errors (RSE) were calculated by dividing the standard deviations by the mean values, and represent the relative accuracy when using a detection ratio for extrapolating to actual crowd size. Also included are the used people counting methods. The automatic video counts during the 'Ghent Festivities 2013' were calculated by a video processing algorithm developed by ViNotion (www.vinotion.nl).

In order to get a better insight in the spatiotemporal variability of Bluetooth detection ratios, we continue by examining the two samples collected by the automatic people counting systems during the 'Ghent Festivities 2013' event. Figure 6.2 on the following page shows the individual detection ratios for each hour. Although the absolute differences are not that large, the Bluetooth detection ratio at the location 'Predikherenlei' seems to follow a distinct pattern over the time of day with slightly lower values early in the morning. The detection ratios at the location 'Jan Breydelstraat' are characterized by a larger amount of scatter, but also show signs of a similar pattern. Additionally, the detection ratio on average — as shown in figure 6.2 on the next page — is also larger at this last location. Further research is necessary into interpreting the differences in average values and patterns between both locations, but it is clear that short-term variations of the Bluetooth detection ratio exist and that future research should take these into account. One important implication is that future crowd size estimations should be based on time-dependent detection ratios instead of one average detection ratio in order to further increase accuracy.

Self-selection bias Besides the general detection ratio, it is also important to investigate the risk of self-selection bias being present in Bluetooth tracking data. Individuals with

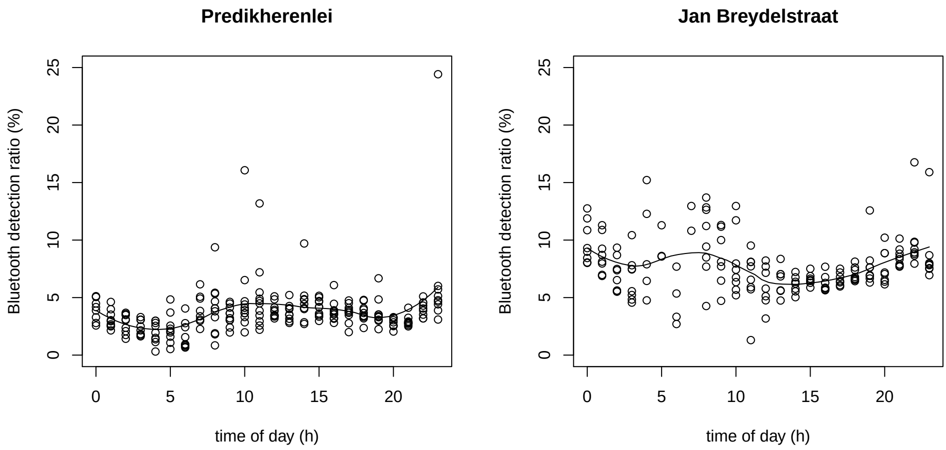


Figure 6.2: Bluetooth detection ratios versus the time of day for two locations during the 'Ghent Festivities 2013' event: 'Predikherenlei' (left) and 'Jan Breydelstraat' (right). Each circle represents the ratio of the number of individuals detected by an automatic people counting system and the number of detected Bluetooth devices over a time period of one hour. The lines represent best-fitting curves through the point clouds based on a LOESS smoother (Cleveland and Devlin, 1988). Data points associated with time periods where less than 50 people were counted by the camera system over one hour were discarded.

certain personal characteristics might, for example, be more likely to carry devices which are detectable and thus be over-represented in the movement datasets. The clear pattern in short-term variations of the Bluetooth detection at the location 'Predikherenlei' shown in figure 6.2 provide a first hint that such a self-selection might be possible due to different audiences passing the sensor location at different moments of the day. In order to gain a direct insight into the degree of self-selection bias, however, we conducted personal interviews during and after the 'Ghent Festivities' event in 2013. This survey sought to investigate the degree in which the detectability of individuals through either Bluetooth or WiFi technology might be related to certain personal characteristics. It must be stressed, however, that this only represents a first attempt in trying to discover and interpret certain trends, and that further research is necessary to validate findings. In total, 243 persons were polled for general personal characteristics (such as age and gender). In parallel, we checked whether the 256 mobile devices they were carrying were detectable through Bluetooth or WiFi technology. After removal of cases where persons opted out during the interview or where detectability could not be determined with certainty, 202 persons with 211 devices remained for Bluetooth and 222 persons with 235 devices for WiFi.

As a first exploration into self-selection bias, we investigated the effect of gender and age. Figure 6.3 on the next page gives a strong indication that gender is indeed related to both Bluetooth and WiFi detectability, where in both cases men are more often detectable. The difference between men and women is somewhat less pronounced for Bluetooth (16% vs. 10%) than for WiFi (27% vs. 16%). As shown in figure 6.4 on the following page, age seems to have an even more pronounced effect on detectability, where Bluetooth-detectable individuals seem mainly older than 30 and WiFi-detectable individuals are generally younger.

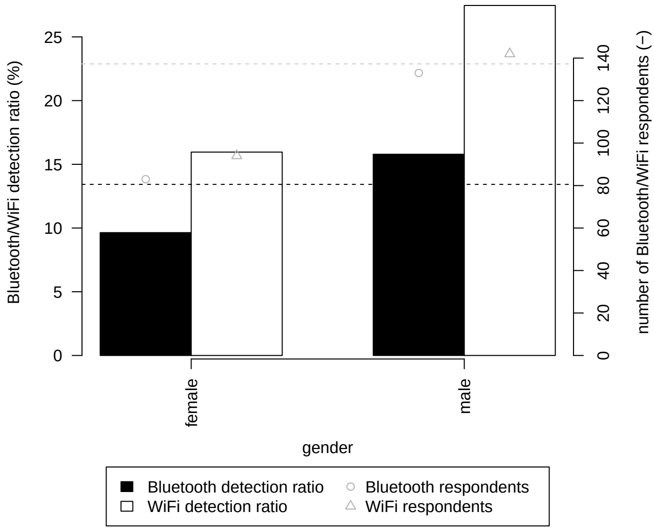


Figure 6.3: Relationship between gender and Bluetooth/WiFi detection ratio. Horizontal dashed lines represent the mean Bluetooth and WiFi detection ratios over all respondents.

Next, we looked at the combined influence of gender and age. The pattern for male respondents in figure 6.5 on page 135 more or less corresponds to the general pattern for age visible in figure 6.4 on the following page. All women over the age of 40 in the survey are not detectable, however. While relatively few women in these age classes were interviewed, this difference is still remarkable and should be further investigated in the future.

Although these preliminary findings suggest significant self-selection bias in both gender and age, it should be clear that they need to be confirmed by future surveys with a better distribution of respondents over both (and other) variables. Because of the rapid evolutions in the distribution of wireless technologies, systematic and repeated surveys should also investigate longer-term temporal variations. In case of Bluetooth, for example, it is not entirely clear whether the technology will continue to spread or diminish. Where alternative technologies such as near field communication (NFC) might cause a downward trend, the Bluetooth Low Energy (BLE) protocol might signal a further growth (although the differences in the detection process between this new protocol and the currently used protocol are not yet clear). Significant spatial variability can also be expected, e.g. between developed and developing countries, similar to other aspects of the 'digital divide' such as

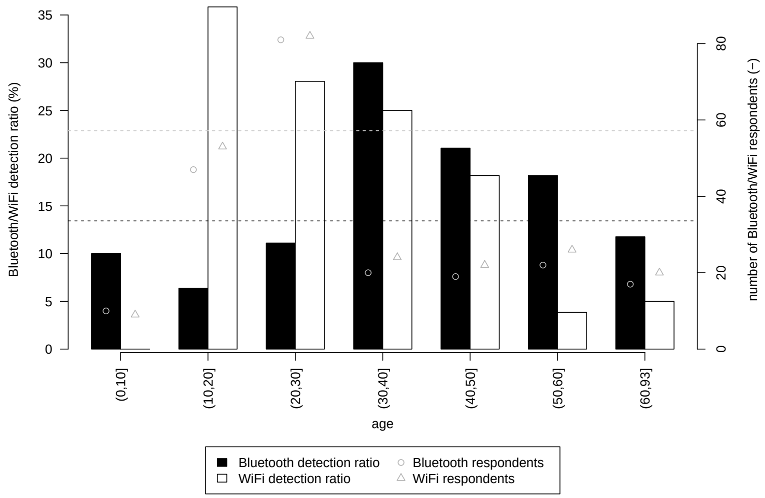


Figure 6.4: Relationship between age and Bluetooth/WiFi detection ratio. Horizontal dashed lines represent the mean Bluetooth and WiFi detection ratios over all respondents.

internet usage (Rice and Katz, 2003; Warf, 2001). This is especially relevant in the context of this specific dissertation as all tracking projects were based in Belgium.

6.2.2.3 Privacy

The aim of this section is to provide a general reflection on location privacy issues, how they apply to Bluetooth tracking, and how our *modus operandi* tries to mitigate most of these risks. Where privacy used to deal with largely static data in the past, there are growing concerns on the rapid proliferation of location-based personal data due to the increasing use of location-aware applications (Beresford and Stajano, 2003; Tsai et al., 2009). As legislators are trying to cope with this rapid development, the legal framework surrounding location privacy is getting increasingly fragmented (Cuijpers and Koops, 2010). Additionally, relatively little attention has gone to the privacy implications of the passive and network-based tracking of mobile devices through ad-hoc sensor deployments (including but not limited to Bluetooth tracking) in comparison to, for example, similar approaches based on data of mobile operators (Green and Smith, 2003) or active handset-based positioning in location-based services (Cuijpers and Pekárek, 2011). As a result, there is no clear legal framework which applies specifically to Bluetooth tracking.

In Belgium, where all tracking projects were based, the legality of the methodology ultimately revolves around tracking data being classified as personal data or not[3]. As Bluetooth MAC addresses cannot be directly linked to identifiable individuals, they are not to

[3]Please note that significant differences may exist between countries, even within the European Union.

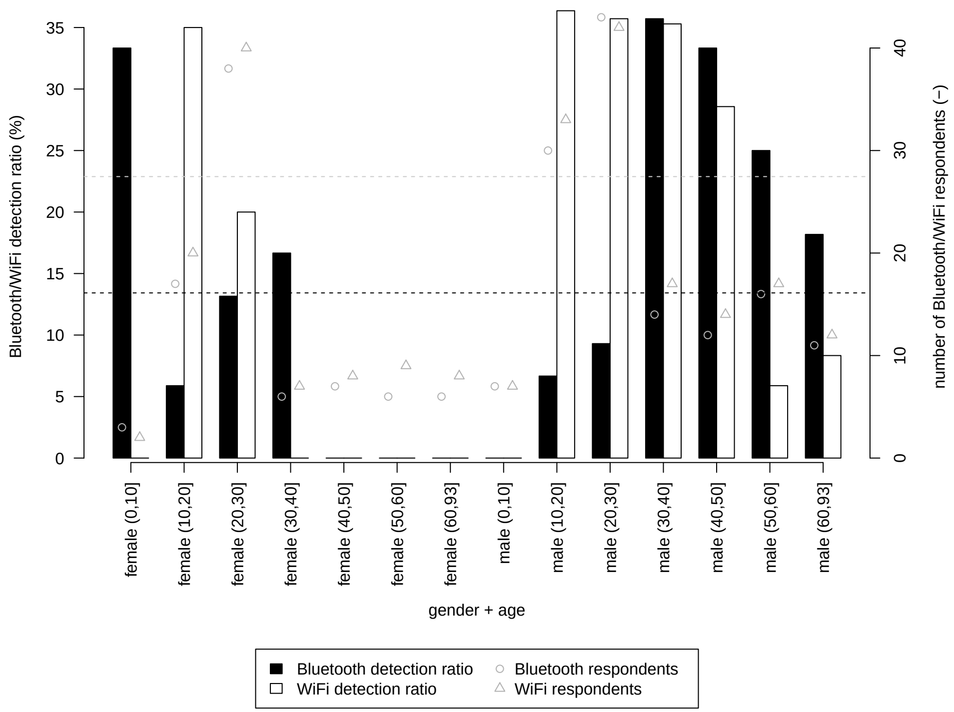


Figure 6.5: Relationship between the combination of age and gender and Bluetooth/WiFi detection ratio. Horizontal dashed lines represent the mean Bluetooth and WiFi detection ratios over all respondents.

be considered as personal data. However, it should be noted that this link could be made indirectly (Kostakos and O'Neill, 2008; Wong and Stajano, 2005). Two approaches are, at least in theory, possible. First, individuals could be visually identified when passing a sensor location thus making it possible to link that individual with the MAC address of his/her mobile device. The low probability of that individual carrying a detectable device, and the labor-intensive nature make this approach less feasible on a larger scale. A second and more comprehensive method is to link MAC addresses with personal information of mobile operators. Technically, this could be done by joining databases of mobile phone manufacturers — linking MAC addresses to IMEI (International Mobile Equipment Identity) numbers — with those of mobile operators — linking IMEI numbers to IMSI (International Mobile Subscriber Identity) numbers which reside on the SIM card. While theoretically possible, this would involve multiple breaches in the security layer of both manufacturers and operators. In sum, these approaches are technically possible but rather unrealistic.

Because Bluetooth tracking is situated in a legal 'grey zone', it is worthwhile to take a closer look at the ethical aspects surrounding the methodology. For instance, several precautionary measures have been taken throughout the last five years in order to further reduce the risk of privacy infringements. First of all, Bluetooth scanners are installed out of reach of the tracked public which reduces the risk of theft and data extraction. Data sent over the internet is encrypted. The friendly names, which could contain personal information, were never recorded nor registered. Although it is impossible to contact every tracked individual to ask for his/her consent, we have always worked under the consent of the event organizers. Additionally, we have always strived for transparency by reaching out to the general audience through press appearances and scientific expositions. Analyses were always performed on an aggregate level. Finally, all projects had a purely scientific purpose. Although this might not seem an important factor in the assessment of privacy risks at first, a survey taken in 2012 (Baeyens, 2012) (N=1,246) clearly revealed a gap in public acceptance of Bluetooth tracking for scientific versus commercial incentives (75% vs. 17% acceptance) and for the application of the methodology for crowd-management or marketing purposes (83% vs. 13% acceptance). This low acceptance for commercial purposes has very recently again been demonstrated by a number of applications based on Bluetooth (and/or WiFi) tracking causing significant public commotion. It only took four days, for example, for an experiment with recycling bins capable of Bluetooth tracking in London to be halted after it was reported by the online magazine Quartz (Datoo, 2013). Future science, technology and society (STS) studies should further investigate the opinions towards new methodologies such as Bluetooth tracking, and how they are possibly influenced by personal factors (age, gender, ethnic background, etc.).

6.3 Conclusions

This dissertation revolved around the study of crowds and their inherently dynamic character. This line of inquiry has traditionally been confronted with a lack of empirical data upon which models and theories could be grounded. In recent years, however, several new methodologies based on the use of mobile phones have been proposed as alternatives for a direct observation of crowd movements. Despite their rather recent introduction, the participatory use of GPS technology and (to a lesser extent) the non-participatory reconstruction of mobility traces through records of mobile phone operators represent two scientifically established methodologies. In contrast, the later proposal of using ad-hoc deployments of sensors based on short-range wireless technologies for the non-participatory tracking of small-scale movements of mobile devices currently lacks a comprehensive academic foundation. By studying the use of Bluetooth technology in tracking crowds (‘Bluetooth tracking’), this dissertation aimed to contribute to building such a foundation.

By investigating and amply illustrating the use of Bluetooth tracking for studying crowds — both through the use of static (Chapter 2) as well as mobile (Chapter 3) sensors — this dissertation set out by focusing on the context of mass events. These contributions allowed readers to get a more comprehensive and tangible view of the methodology in comparison with the slim body of literature to date. A special focus was also given to the spatiotemporally sparse or 'episodic' nature of Bluetooth tracking data trajectories. The model developed in Chapter 4, which was used for calculating the potential movement of a crowd in between sensor locations, represented a first attempt at adapting traditional paradigms and methods which usually consider trajectories with fixed sampling frequencies. As such, the model's application is not limited to Bluetooth tracking data alone but spans the entire range of methodologies which gather similarly structured movement data. The case study on tourist behavior in Chapter 5 sought to demonstrate that Bluetooth tracking is not limited to studying crowds at mass events, but can also be used for studying more general groups of individuals with similar intentions. The lack of explaining socio-economic or psychographic variables in the tracking data did, however, necessitate a shift from a hypothesis-based analysis to a data-mining approach with no a-priori assumptions. Despite the short tracking period and the need for an extensive filtering procedure, the association rule learning scheme was able to discover interesting patterns which are potentially valuable for improving tourism management practices. Appendix A was finally dedicated to the *GISMO* toolkit which was developed for analyzing the Bluetooth tracking data which were gathered in the different case studies. Besides playing an instrumental role in the realization of this dissertation, the toolkit also represents the first comprehensive software implementation dedicated to the analysis of episodic movement data.

The ensuing discussion reflected on a selection of important issues which require further attention in the future. A deeper interpretation of crowd behavior will first of all necessitate the combination of Bluetooth tracking with approaches polling for personal backgrounds and motivations of tracked individuals. How this combination should be realized is still unclear, however. A closer look at the detection ratio then demonstrated that the share of individuals detectable through a Bluetooth device seems to be decreasing. To cope with this issue, future tracking approaches should probably include other wireless technologies than Bluetooth alone (most notably WiFi technology). Preliminary results from a survey performed in the summer of 2013 additionally indicated that the detectability through both Bluetooth and WiFi technology seems to be related to personal characteristics such as gender and age. This resulted in a self-selection bias where certain groups are over- or under-represented. More specifically, both technologies seemed to exhibit a bias of men over women, Bluetooth technology seemed to reach an older crowd than WiFi technology, and women over the age of 40 seemed to be missing. Further studies should take a closer look at this self-selection bias and propose ways of mitigating its effect.

References

Abowd, G. D. and Mynatt, E. D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction*, 7(1):29–58.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., and Hecker, D. (2012). Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. *KI - Künstliche Intelligenz*, 26(3):241–251.

Atzori, L., Iera, A., and Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805.

Baeyens, T. (2012). *Bluetoothtracking : Privacy en representativiteit.* Msc.

Beresford, A. and Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55.

Blondel, V., Krings, G., and Thomas, I. (2010). Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42:1–12.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596.

Cuijpers, C. and Koops, B.-J. (2010). How fragmentation in European law undermines consumer protection: the case of location-based services. *European Law Review*, 33:880–897.

Cuijpers, C. and Pekárek, M. (2011). The regulation of location-based services: challenges to the European Union data protection regime. *Journal of Location Based Services*, 5(3-4):223–241.

Datoo, S. (2013). This recycling bin is following you.

Eagle, N. and Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

Fang, Z., Tu, W., Li, Q., and Li, Q. (2011). A multi-objective approach to scheduling joint participation with variable space and time preferences and opportunities. *Journal of Transport Geography*, 19(4):623–634.

Farber, S., Neutens, T., Miller, H. J., and Li, X. (2013). The Social Interaction Potential of Metropolitan Regions: A Time-Geographic Measurement Approach Using Joint Accessibility. *Annals of the Association of American Geographers*, 103(3):483–504.

Fielding, N. and Schreier, M. (2001). Introduction: On the Compatibility between Qualitative and Quantitative Research Methods. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 2(1):1–14.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Green, N. and Smith, S. (2003). 'A Spy in your Pocket'? The Regulation of Mobile Data in the UK. *Surveillance & Society*, 1(4):573–587.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational researcher*, 17(8):10–16.

Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J., and Diot, C. (2005). Pocket switched networks and human mobility in conference environments. In *Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking (WDTN '05)*, pages 244–251, New York, NY. ACM Press.

Kang, H. and Scott, D. M. (2007). An integrated spatio-temporal GIS toolkit for exploring intra-household interactions. *Transportation*, 35(2):253–268.

Kostakos, V. and O'Neill, E. (2008). Capturing and visualising Bluetooth encounters. In *Adjunct proceedings of the Conference on Human Factors in Computing Systems (CHI 2008)*, Florence.

Larsen, J. E., Sapiezynski, P., Stopczynski, A., Moerup, M., and Theodorsen, R. (2013). Crowds, Bluetooth, and Rock-n-Roll. Understanding Music Festival Participant Behavior.

Leitinger, S., Gröchenig, S., Pavelka, S., and Wimmer, M. (2010). Erfassung von Personenströmen mit der Bluetooth-Tracking- Technologie. In *Angewandte Geoinformatik 2010*, pages 220–225, Salzburg, Austria.

Miller, H. (2005). Necessary space - time conditions for human interaction. *Environment and Planning B: Planning and Design*, 32(3):381–401.

Millonig, A. and Gartner, G. (2007). Monitoring Pedestrian Spatio-Temporal Behaviour. In Gottfried, B., Van de Weghe, N., Billen, R., and De Maeyer, P., editors, *Workshop on Behaviour Monitoring and Interpretation (BMI '07)*, pages 29–42, Ghent.

Neutens, T., Farber, S., Delafontaine, M., and Boussauw, K. (2013). Spatial variation in the potential for social interaction: A case study in Flanders (Belgium). *Computers, Environment and Urban Systems*, 41:318–331.

Neutens, T., Schwanen, T., Witlox, F., and De Maeyer, P. (2008). My space or your space? Towards a measure of joint accessibility. *Computers, Environment and Urban Systems*, 32(5):331–342.

Neutens, T., Versichele, M., and Schwanen, T. (2010). Arranging place and time: A GIS toolkit to assess person-based accessibility of urban opportunities. *Applied Geography*, 30(4):561–575.

Neutens, T., Witlox, F., and Demaeyer, P. (2007). Individual accessibility and travel literature review on time geography possibilities : A literature review on time geography. *European Journal of Transport Infrastructure and Research*, 7(4):335–352.

Rice, R. E. and Katz, J. E. (2003). Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts. *Telecommunications Policy*, 27(8-9):597–623.

Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., and Oliveira, A. (2011). Smart Cities and the Future Internet : Towards Cooperation Frameworks for Open Innovation. In Domingue, J., Galis, A., Gavras, A., Zahariadis, T., Lambert, D., Cleary, F., Daras, P., Krco, S., Müller, H., Li, M.-S., Schaffers, H., Lotz, V., Alvarez, F., Stiller, B., Karnouskos, S., Avessta, S., and Nilsson, M., editors, *The Future Internet - Future Internet Assembly 2011: Achievements and Technological Promises*, volume 6656 of *Lecture Notes in Computer Science*, pages 431–446. Springer, Berlin, Heidelberg.

Shoval, N. and Isaacson, M. (2009). *Tourist mobility and advanced tracking technologies*, volume 19 of *Routledge Advances in Tourism.* Routledge, New York, London.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Stopczynski, A., Larsen, J. E., Lehmann, S., Dynowski, L., and Fuentes, M. (2013). Participatory Bluetooth Sensing: A Method for Acquiring Spatio-Temporal Data about Participant Mobility and Interactions at Large Scale Events. In *International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications 2013*, pages 242–247, San Diego, CA.

Tsai, J. Y., Kelley, P. G., Cranor, L. F., and Sadeh, N. (2009). Location-Sharing Technologies: Privacy Risks and Controls. In *37th Research Conference on Communication, Information and Internet Policy (TPRC 2009)*, Arlington, Virginia.

Warf, B. (2001). Segueways into cyberspace: multiple geographies of the digital divide. *Environment and Planning B: Planning and Design*, 28(1):3–19.

Watson, R. and Yip, P. (2011). How many were there when it mattered ? *Significance*, 8(3):104–107.

Wong, F. and Stajano, F. (2005). Location privacy in bluetooth. In *Security and privacy in ad-hoc and sensor networks*, pages 176–188. Springer.

A

GISMO: a Geographical Information System for the analysis of Moving Objects based on episodic proximity-based sensor tracking data

Versichele, M., Neutens, T., Moerman, I., Van de Weghe, N.

Abstract *In this chapter, we advocate the need for further attention to episodic movement data. We first focus on the spatiotemporal sparseness of this movement data type, and demonstrate its growing importance in empirical mobility studies. More specifically, we focus on tracking methodologies that use ad-hoc sensor networks for proximity-based location registrations of mobile devices. A toolkit — called* GISMO *— for the analysis of such data is presented, and an overview of its main capabilities (preprocessing, transformation, selection, visualization) is given by a case study using a tracking dataset. Although the software has been originally conceived for Bluetooth tracking data, any kind of proximity-based dataset can be used as long as it is correctly formatted.*

A.1 Introduction

Human mobility on different spatial and temporal scales plays a pivotal role in many processes taking place in our environment. The increasing inter-connectedness of our globalized world poses significant challenges in different contexts such as the sustainable organization of urban expansion (Camagni et al., 2002) or the mitigation of global epidemics (Hufnagel et al., 2004). The traditional way of measuring human movement through direct visual observation or questionnaires is ill suited for such large-scale processes, but objects that are somehow associated with or carried around by persons are often easier to measure. Several of these ‘proxies’ for human movement have already been studied, including one-dollar bank notes (Brockmann et al., 2006) and public transit smart cards (Pelletier et al., 2011). Tourist movements, in particular, are sometimes studied by distributing GPS (global positioning system) logging devices for the duration of their visit (Shoval and Isaacson, 2007). It is the use of mobile phones, however, that is starting to truly revolutionize the field as they are equipped with a wide variety of (location-aware) technologies, and are often in close contact with their owners.

Digital traces left behind by mobile phone users that can be mined for interesting knowledge can either be of a voluntary or involuntary nature. Voluntary traces include geo-tagged pictures (Girardin and Calabrese, 2008; Jankowski et al., 2010) and data from location sharing services such as Foursquare (Cheng et al., 2011). Involuntary traces can be contained, for instance, in call logs of mobile phone operators. In contrast to voluntary traces, such datasets are not affected by a potentially large variability in user commitment. As a result, a growing number of studies has utilized these types of datasets (Ahas et al., 2008; Candia et al., 2008; González et al., 2008). In order to avoid any potential privacy infringements, these datasets have always first undergone a thorough anonymization process. The potentially complicated cooperation with mobile operators and the limited positional accuracy for smaller-scale studies (Ahas et al., 2007) has, however, caused some researchers to adopt an alternative approach in which an ad-hoc network of sensors at strategic locations is used to measure movements in between these locations.

Bluetooth, in particular, has already been used as such an ad-hoc sensing technology for various purposes including mass-event crowd analyses (Delafontaine et al., 2012; Larsen et al., 2013; Stange et al., 2011; Versichele et al., 2012a,b), travel time measurements along highways (Haghani et al., 2009), transit time measurements in airports (Hainen et al., 2013), and urban design studies (O’Neill et al., 2006). Some of the advantages of the technology include its ubiquitous applicability (indoor and outdoor, all weather and lighting conditions, etc.), its ability to track movements without the direct involvement of the tracked individual, and the relatively low cost of the necessary hardware (Versichele et al., 2012a). Recently, a similar ad-hoc and unobtrusive methodology but based on WiFi technology has also been proposed (Bonné et al., 2013), while large public local area networks can also be mined for digital traces left behind by mobile devices connecting to several access points (Ojala et al., 2005). In parallel to these trends, considerable research efforts have also gone to indoor localization methodologies using WiFi (Mazuelas et al., 2009; Vanheel et al., 2011, 2013) or Bluetooth (Zhou and Pollard, 2006) signal strengths. In the remainder of this chapter, however, we will focus on a system where signal strengths are not used for localization.

Regardless of the differences in technology or deployment, the previously mentioned network-based tracking methodologies (either through an existing or ad-hoc infrastructure) generate a type of movement data with significantly different characteristics compared to more common GPS-based trajectories. Most importantly, the tracking data is spatiotemporally sparse. In the spatial dimension this is evident for ad-hoc deployments as trajectories are constrained to moves in between sensor locations. Despite the larger coverage of existing mobile phone infrastructure, the recorded locations are usually also restricted to the exact locations of the cell-towers the phone is connected to. For ad-hoc deployments, this proximity-based positioning (Bensky, 2007) automatically causes an additional temporal sparseness as devices that are in between sensor ranges are not detected. Mobile phone records are usually further constrained by (cell-tower) locations only being registered when performing an activity such as making a phone call or sending a text message. The growing importance of such data, which are known in literature as episodic movement data (Andrienko et al., 2012), constitutes a challenge for traditional methods or toolkits as they are usually built on the notion of spatiotemporally continuous GPS-like trajectories. Additionally, the data type is not limited to network-based tracking methodologies as such, because GPS trajectories are sometimes generalized (e.g. for handling privacy issues) leading to a data structure which is in essence also episodic (Monreale and Andrienko, 2010).

This chapter does not seek to argue that the analysis of episodic movement data is an insurmountable challenge. Rather, we do believe that that there currently is a shortage of user-friendly software for analyzing this specific data type. While episodic movement data bear some resemblances to space-time paths as conceptualized in the field of time geography (Hägerstrand, 1970), these paths are only sparse in the temporal dimension. Additionally, toolkits analyzing these paths usually focus on the calculation of space-time accessibility measures (Neutens et al., 2010; Shaw and Yu, 2009; Shaw et al., 2008; Yu and Shaw, 2008) and were not specifically tailored to perform any other tasks. Visual analytics and data mining methods, on the other hand, have already been applied to episodic movement data (Andrienko et al., 2012; Stange et al., 2011), but they have yet to be collectively implemented in one toolkit and are usually restricted to the steps of 'data mining' and 'interpretation' and 'evaluation' as represented in the traditional knowledge discovery in databases (KDD) process (Fayyad et al., 1996).

The aim of this chapter is to present a toolkit — named *GISMO* — that is specifically designed to analyze episodic movement data, offering all commonly used steps in the KDD process including lower-level tasks, such as preprocessing, selection and transformation, but also visualization options under one user-friendly graphical user interface (GUI). Data mining tasks are not part of its scope, but the output of the toolkit can be used by the many data-mining options available right now such as *WEKA* (Witten et al., 2010), *Orange* (Demšar et al., 2004), or *R* (Ihaka and Gentleman, 1996). The presented toolkit was developed for use with Bluetooth tracking data, but in principle any episodic dataset can be used as long as it is in the correct format. We will illustrate the toolkit's main capabilities by analyzing a Bluetooth tracking dataset gathered at a music festival.

The remainder of the chapter first presents some more details on the Bluetooth tracking methodology and the Bluetooth tracking dataset in section A.2. The overview of the toolkit in section A.3 on the facing page starts with an explanation of its graphical user interface in section A.3.1 on the next page. Subsequently, we analyze the data in section A.3.2 on page 149. The chapter concludes with some final remarks in section A.4 on page 159, focusing on the merits of the toolkit as well as on its limitations.

A.2 Bluetooth tracking methodology and dataset

Bluetooth tracking is a methodology in which embedded computers equipped with Bluetooth sensors are used to track devices with a visible Bluetooth interface. The strategic placement of these sensors can generate informative trajectories, which can be analyzed for different purposes. In brief, the discovery process starts with an inquiry by the sensor, which is responded to by devices within the detection range that have their Bluetooth interface set to 'discoverable'. The inquiries (which by default take 10.24 seconds) are continuously repeated, and all responses are logged by the computer to which the sensor is connected. Because devices are repeatedly detected as long as they are in the detection range, detections are converted to detection intervals on the fly by using a buffer time under which subsequent detections of the same device are converted to a time interval. The buffer time is also set to 10.24 seconds. More details on the methodology and this conversion process can be found in (Versichele et al., 2012a).

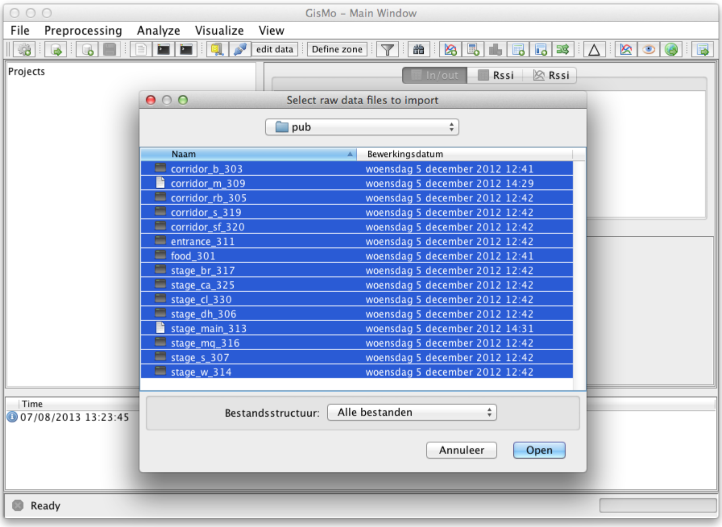
The resulting dataset consists of log files containing episodic detection data with the following comma-separated format: *detection timestamp, MAC address, class of device, 'in'/'out'/ 'pass'*. The *MAC* address represents a 48-bit unique hardware identifier of the mobile device. The first 24 bits identify the manufacturer of the device, so we can deduce the brand of the detected devices. The class of device code can be used to deduce the type of device and its services according to the Bluetooth protocol. Finally, 'in' represents the start of a detection episode, 'out' the end of a detection episode, and 'pass' a solitary detection with no previous or subsequent detections within 10.24 seconds. The dataset that will be used to demonstrate the *GISMO* toolkit was gathered at a three-day music festival in 2012 that attracted over 60,000 visitors per day. Fifteen locations (eight stages, five corridors between the stages, the entrance and the backstage restaurant for the festival crew) were covered by a Bluetooth sensor (see figure A.16 on page 158 for a reference). The fifteen log files together contained 2,802,336 log lines.

A.3 Overview of the *GISMO* toolkit

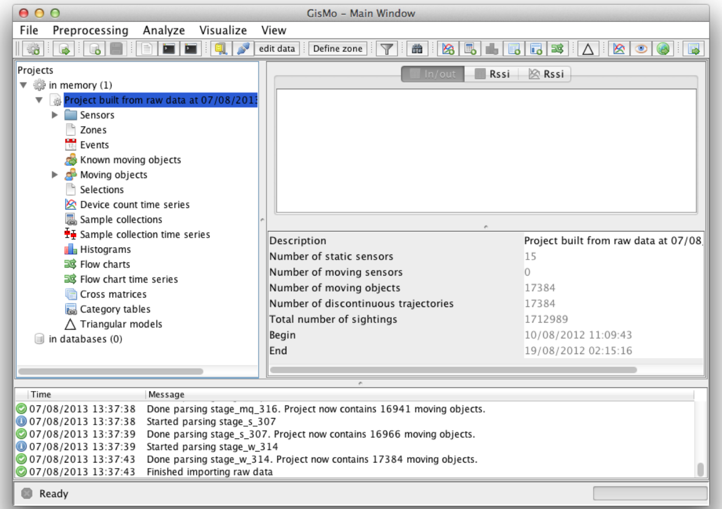
In this section, we will present an overview of the *GISMO* toolkit. The toolkit was developed in the *Java* programming language, making it cross-platform. All analyses were done on a non-dedicated system with a 2.66 Ghz Intel Core 2 Duo CPU and 8 Gb of internal memory. The operating system was Mac OS X 10.7.5, and the JRE (*Java* Runtime Environment) was *Java* SE 6. We will begin by explaining the toolkit's graphical user interface (GUI), and continue by outlining some of its capabilities by importing, preprocessing and analyzing the Bluetooth tracking dataset presented in section A.2 on the preceding page. For the sake of clarity, we will format terminology related to the toolkit in italics. Labels depicted in the figures are surrounded by single brackets. The graphical user interface of *GISMO* consists of three separate panels and a menu- and toolbar on top, as shown in figure A.1 on the following page and figure A.2 on page 147. The lower panel is a status panel giving indications of the program's progress. The left panel contains a tree structure labeled the 'Projects' tree, which contains all projects the user is working on. Each project consists of some (core) data objects, such as sensors, detected Bluetooth devices and trajectories. Additionally, the user can generate metadata which represents the output of certain algorithms applied to the data (section A.3.2 on page 149). The right panel is subdivided into two smaller subpanels. The upper subpanel reveals a tabular structure when selecting appropriate objects in the 'Projects' tree. The lower subpanel lists more general and aggregated information on the selected objects. We will now import and preprocess the Bluetooth tracking dataset.

A.3.1 Importing and preprocessing

The log files of the dataset — formatted as described in section A.2 on the preceding page — can be directly imported in the *GISMO* environment, as shown in figure A.1a on the following page. Once the import is done, the different GUI components are populated with the information linked to the selection made in the 'Projects' tree. Figure A.1b on the next page shows which information is listed when the newly created project is selected. We can see that the dataset consists of 17,384 unique Bluetooth devices that were detected over 1,712,989 times ('sightings') by the fifteen deployed sensors, and that the data spans ten days. After renaming the project (in our case to 'Festival') and drilling down the 'Sensors' folder, the GUI shows information linked to the selected sensor (figure A.2a on page 147). The sensor at the entrance, for example, detected 8,172 devices over 128,677 intervals and was set up on the 15th of August. Additionally, the geographical coordinates of all sensors can be set for further visualization purposes. Further down the tree, there is a folder labeled 'Moving objects' that contains a node for each detected Bluetooth device in the dataset. Selecting such a device lists the time-ordered detections associated with it in the upper right panel, as shown in figure A.2b on page 147. The lower panel reveals that the selected device is a Nokia smartphone that was first detected at the entrance and subsequently over 12 locations in total over five days. The device service classes seem to suggest that the phone



(a)

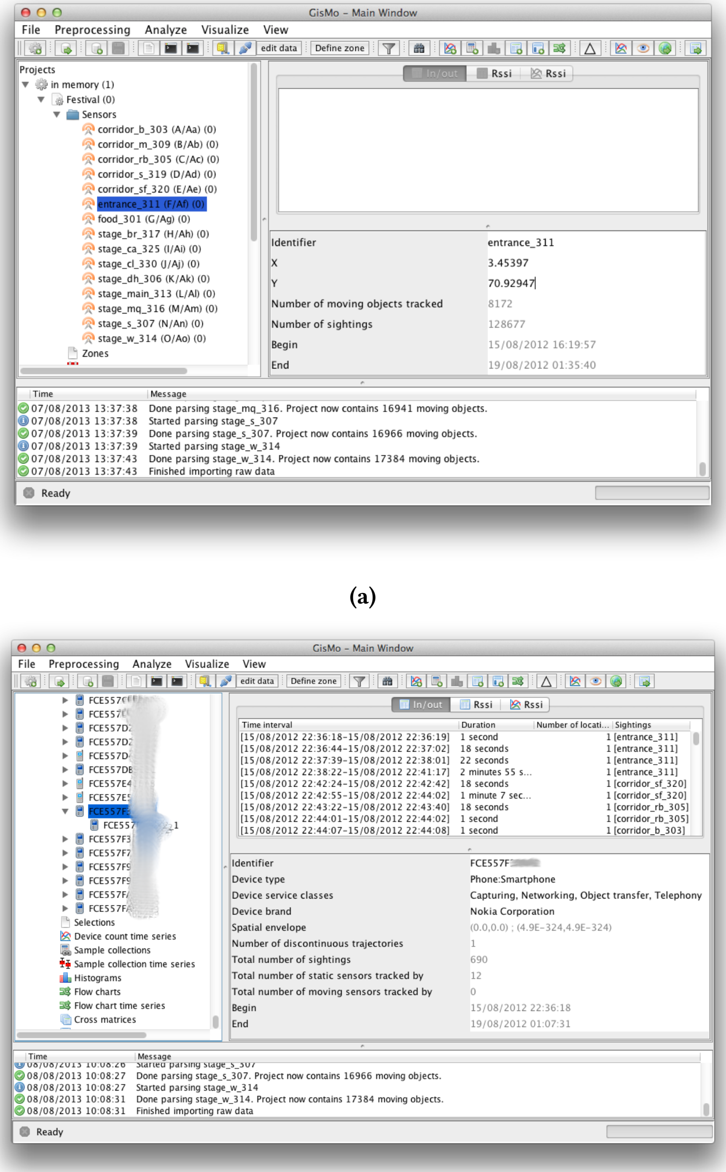


(b)

Figure A.1: Overview of the *GISMO* graphical user interface showing the import of the raw log files (a), and the project properties (b).

has a camera (indicated by the 'capturing' service).

At this stage, each device is associated with one trajectory depicted by a child node of the device node in the 'Projects' tree and carrying the name of the device's MAC address



(a)

(b)

Figure A.2: Overview of the *GISMO* graphical user interface showing the sensor properties (a) and the detections associated with one Bluetooth device after import (b).

followed by '_1'. This trajectory contains all detections of the device. For some analyses, it makes more sense to split trajectories when a long time passes between two subsequent detections. In the context of the festival, this could be between a last detection at the entrance upon exiting the festival area after midnight and a subsequent detection at the entrance upon entering the festival area the next day. In *GISMO*, trajectories can be split using a *maximum gap in seconds* parameter. This way, trajectories will be split at all gaps with a duration longer than this parameter. The process and its effect is visible in figure A.3, where the selected device is now associated with three trajectories after setting the parameter to 18,000 seconds (5 hours).

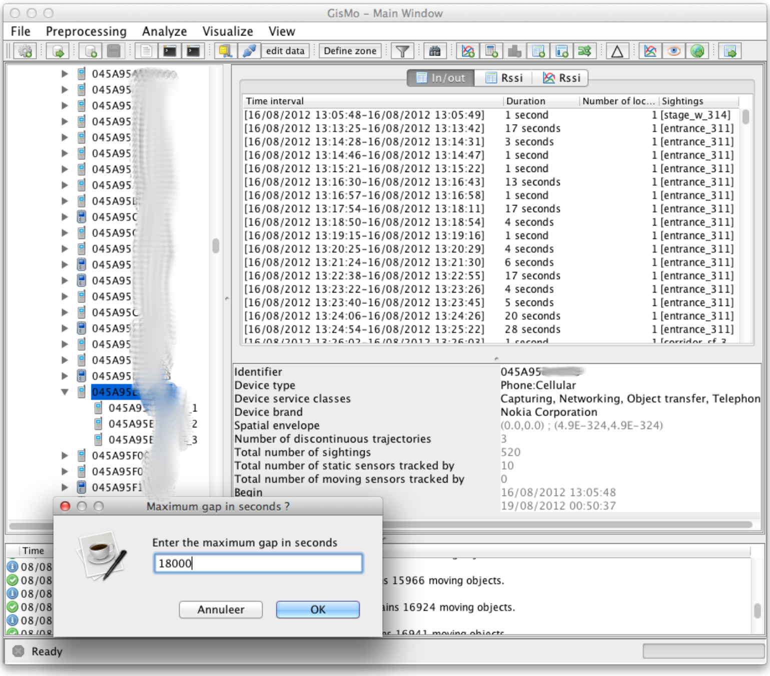


Figure A.3: Splitting of trajectories according to the *maximum gap in seconds* parameter.

As can be seen in figure A.1a on page 146, the project is currently classified under the 'in memory' branch. This means that all operations on the data are lost when the program quits. In order to avoid this, we now save the project in a database format[1]. After exporting the project to a database format, it is classified under the 'in databases' branch as can be seen in figure A.4 on the facing page. We are now ready to start analyzing the data.

[1]This is actually not a database *sensu stricto*, but a directory structure that saves all data and metadata. Despite this technical difference, the possibility to save the progress in processing and analyzing a dataset feels like working on a database to the user for all intents and purposes.

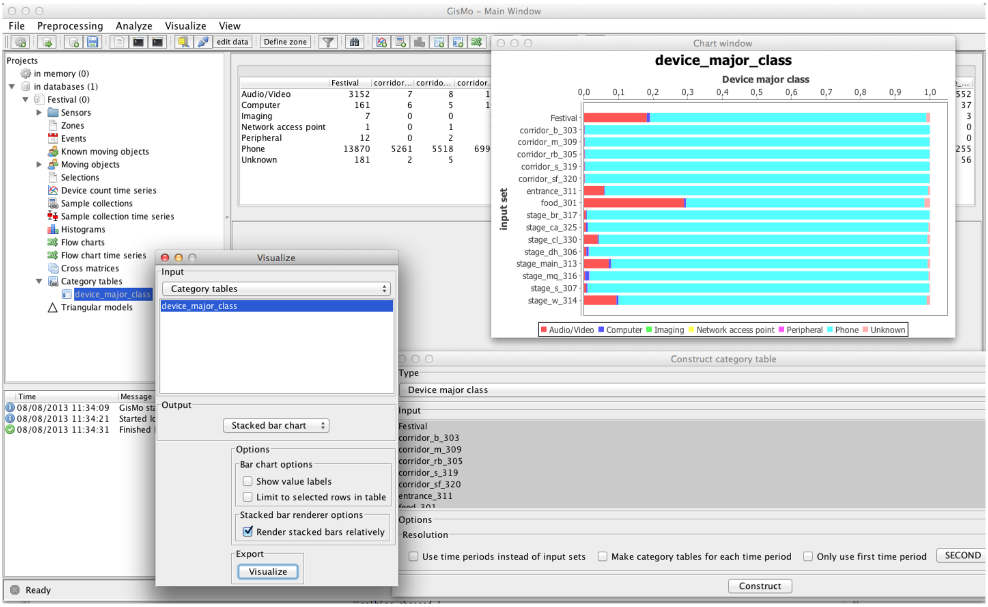


Figure A.4: Calculation and visualization of a *category table* containing the number of devices from each major class at the different locations and the entire project.

A.3.2 Analyses

As described in section A.2 on page 144, each detection is accompanied by the class of the detected device. As it is our intention to explore pedestrian movements, we need to investigate whether the detected devices are appropriate as proxies. This is not a trivial question, as a wide variety of devices is nowadays equipped with Bluetooth technology and some of these devices are either stationary (e.g. desktop computers) or linked to vehicles (e.g. carkits). Exploring the types of Bluetooth devices detected at different locations can be done by calculating a *category table* in *GISMO*, the dialog for which is visible in the lower left corner of figure A.4. After specifying a name, the result is added under the 'Category tables' branch. A tabular representation is automatically shown in the GUI when a user selects the node, a visual representation can be plotted through the 'Visualize' dialog. In this case, a stacked bar chart has been plotted showing that all corridors and most stages almost exclusively detect phones (Major class = 'Phone'). Only the entrance, three stages and especially the backstage restaurant are associated with significant shares of 'Audio/Video' devices. Upon further inspection, these devices belong to the minor classes 'Handsfree' or 'Headset' indicating that these devices are often associated with vehicles. Not surprisingly, the five locations with higher shares of these devices are all located in close proximity to the road to the west of the festival area (figure A.15 on page 157). Because we are now working in a database project, the disk icon in the toolbar (fourth from left) has become enabled indicating the metadata we just created can be saved to the database.

In order to further investigate the differences between the (major) device classes, we first make a selection for each class as shown in figure A.5. This is done in the 'Make new selection' dialog, where an *attribute constraint* (constraining on the major class, minor class, device brand, or service class) is used. Other possible constraints include *spatial* ('device was detected by sensor A'), *temporal* ('device was detected at t'), *spatiotemporal* ('device was detected by sensor A at t'), *sequential* (using regular expressions of sensor characters), *selectional* ('device is part of selection S'), and *sampled property value* (cfr. sample collections). All of these constraints can be combined through an *and/or* operator. Because 'Make new selection' is chosen, the new selection will appear under the 'Selections' branch of the 'Projects' tree. This branch can then be browsed just like the 'Moving objects' branch containing all devices in the dataset. The GUI additionally reports the number of devices ('moving objects') and trajectories that were found.

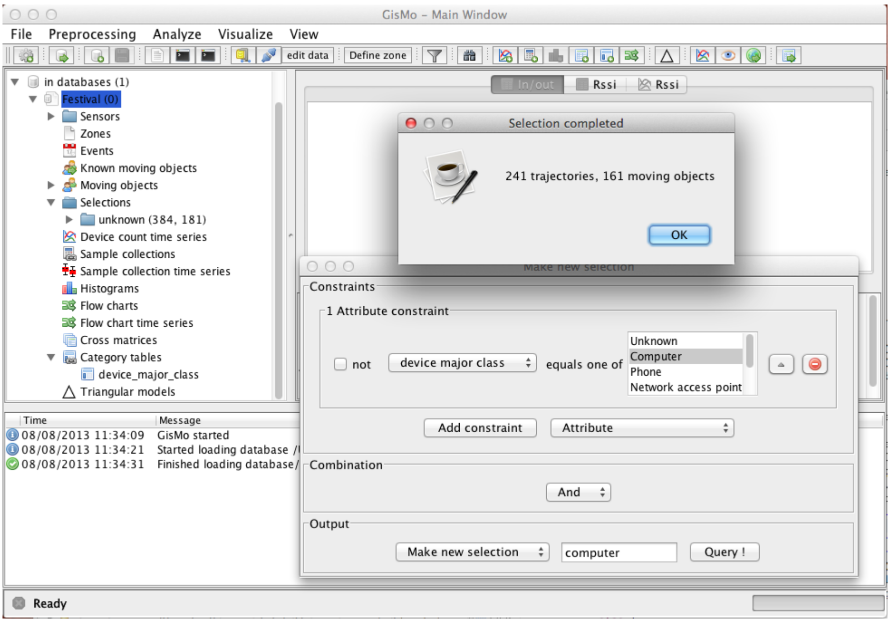


Figure A.5: Calculation of a new *selection* which only contains Bluetooth devices of the 'Computer' major class.

To investigate whether the phones show a different spatial distribution than other devices (which may be static in nature and thus only linked to one location), we can take a *sample* of each selection that was previously made, as depicted in figure A.6 on the next page. Out of all *samplable properties*, we first select the one(s) that we want to calculate. We then limit to each appropriate selection in order to only sample the trajectories in that selection. Each time, a *sample collection* is added to the data tree. The 'collection' terminology was chosen because a *sample collection* can consist of multiple samples, each sample associated with a sampled property (e.g. each device can be sampled for the number of locations it was detected at and the time difference between the first and last detection). Selecting a sample in the GUI reveals the individual sampled objects and property values in the top-right table. Some statistical summary statistics are listed beneath this table. The median value of 1 visited location for the devices with major class 'Unknown' already seems to suggest that these devices are not representative for capturing visitor movements, but a histogram for all device classes might offer a better overview.

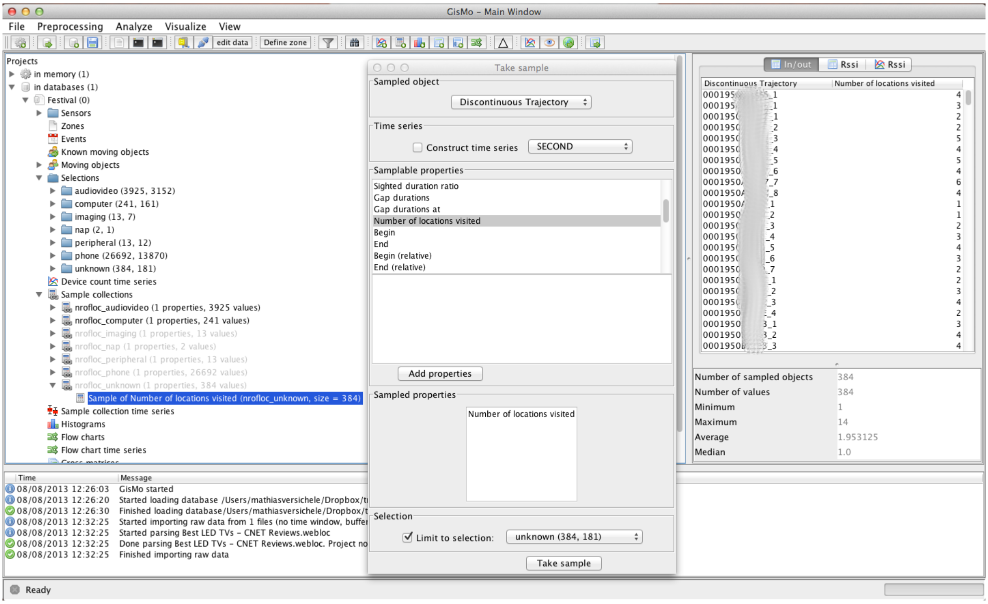


Figure A.6: Calculating a *sample* of the 'Number of locations visited' property for each previously calculated selection based on the major device class.

Whenever a *sample (collection)* is selected in the 'Projects' tree, the icon for creating a *histogram* is enabled. When it is clicked, a dialog is opened where the bin width and the minimum threshold need to be specified. Figure A.7 on the following page shows this dialog and the first histogram created for the *sample collection* calculated for the Audio/Video devices, now lodged under the 'Histograms' branch. On the right, we can see the bins, and the absolute and relative frequencies of property values within the limits of each bin.

After all seven histograms are calculated, we can plot them simultaneously through the 'Visualize' dialog (the same dialog as shown in figure A.4 on page 149). The result is visible in figure A.8 on page 153. Phones clearly have the largest spatial spread of all the classes, Audio/Video devices the smallest. The other classes have a much smaller footprint in the data. In the remainder of the analyses, we will constrain to phones.

The number of devices detected by a sensor is representative for the crowdedness of its

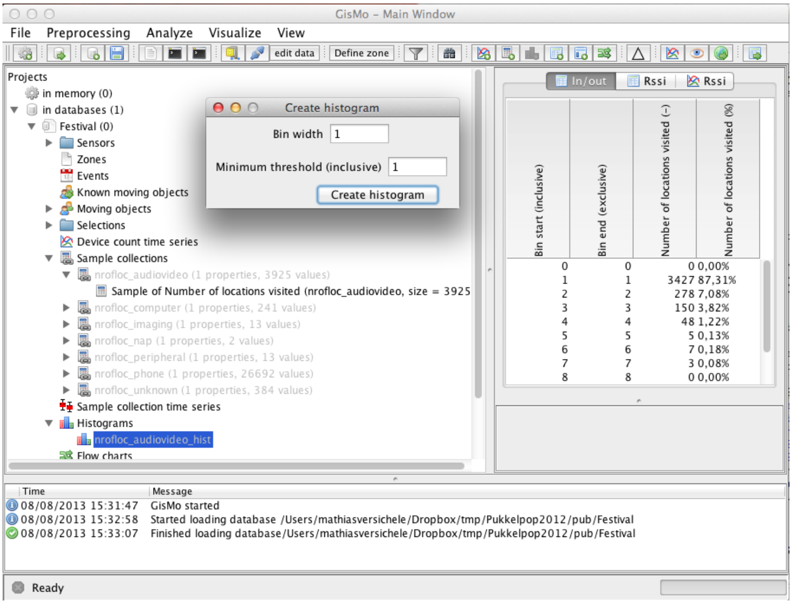


Figure A.7: Calculation of a *histogram* based on a selected *sample*.

location. By calculating a time-series, we can get an insight into the temporal nature of the crowdedness. In *GISMO*, we can do this by calculating another type of metadata: a *device count time series*. Figure A.9 on the next page shows the dialog for doing so in the lower left corner, together with two plots. In the input dialog, the user needs to specify the locations for which time-series need to be generated together with the temporal resolution. Additionally, we also specified that only phones need to be counted by checking the 'Limit to selection' checkbox. Just like the other metadata, the calculation results appear in the 'Projects' tree and can be examined in a tabular fashion or plotted graphically. The graph in the top-right corner shows the time-series for all eight stages over the three festival days. The graph beneath shows the time-series for one individual stage and is further annotated with the bands playing that day. These events which are also listed in the 'Projects' tree were manually entered in a text format in the project database folder.

Before continuing the analysis, we should investigate how long devices are generally detected when someone passes a sensor location. We again take a *sample collection*, now based on the 'staying time' property at each location. As figure A.10 on page 154 shows, selecting this *samplable property* automatically populates the table below with the staying time at each location. We select all these properties, and constrict to phones. The new sample collection now consists of 15 samples. We calculate a histogram, as was previously shown in figure A.7 and visualize it. The resulting plot is also shown in figure A.10 on page 154. It becomes clear that all locations are associated with a very high share of solitary detections (intervals with a duration of 1 second, and no detections prior or later within 10.24

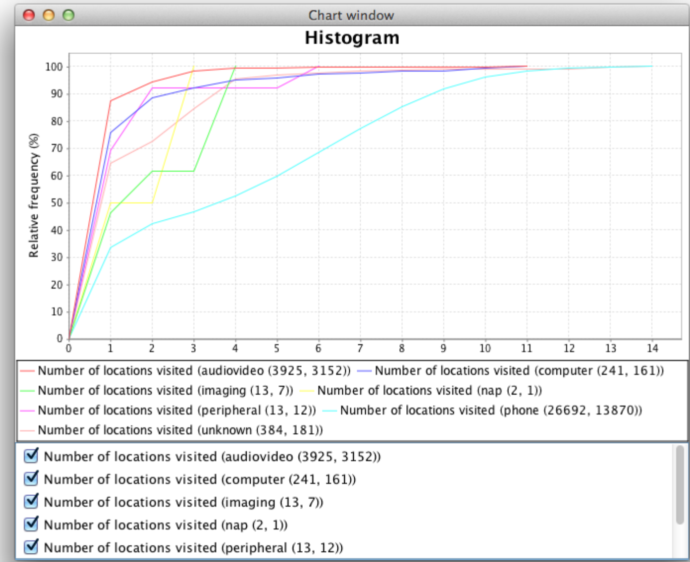


Figure A.8: Plot showing the relative cumulative distribution functions for the 'number of locations visited' property, subdivided over all major classes.

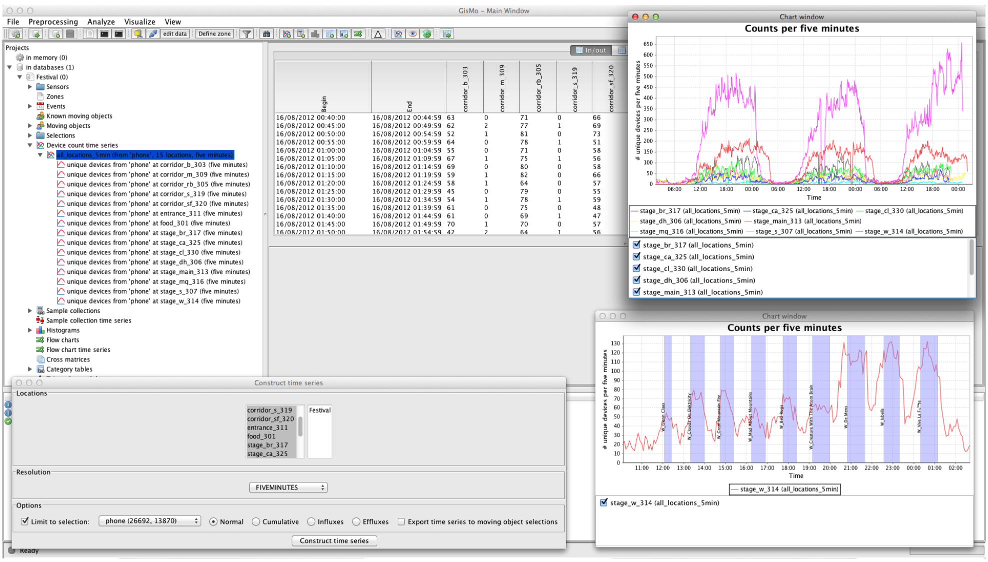


Figure A.9: Calculation of a *device count time series.*

seconds). These very weak and almost coincidental detections are caused by devices further away from the sensor. Once a device gets closer, detections follow up each other more closely leading to actual detection *intervals*. The abnormal frequency of these instantaneous detections is an indication that these may be considered as noise in the data, and should be filtered out.

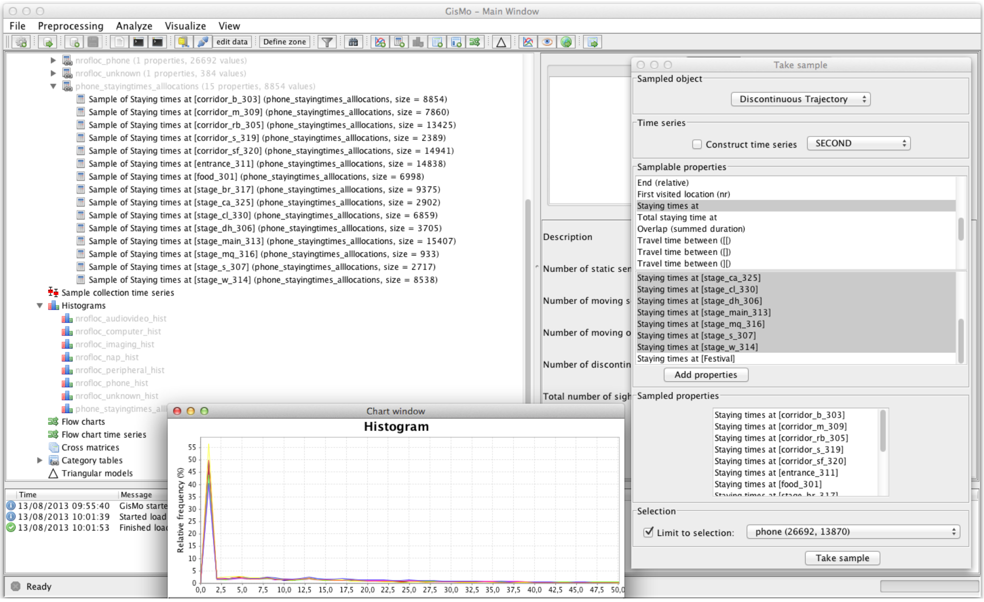


Figure A.10: Calculation of a *sample collection* of the durations of detection ('staying time') at each sensor location and visualization of the histogram.

Filtering in *GISMO* is done through the 'Create filtered view' dialog, which is shown in figure A.11 on the next page. In this case we choose to 'remove sightings strictly smaller in length than' 2 seconds. There are many more filtering options available, one of which is the 'compression' of detection intervals. With this option, subsequent and co-located detection intervals which are separated by a time gap shorter than the specified threshold can be compressed into one longer interval. The effect of a filter combining both options is shown in figure A.12 on page 156. This filter is a *live* filter, meaning that it applies throughout the GUI and any further calculations as long as it is not turned off by unchecking the first checkbox in the dialog. As soon as the filter is turned off, all data returns to its original state.

Now that we know how a filter works, we will investigate the travel time between the entrance and the first corridor behind the entrance ('corridor_sf'). If we define this travel time as the time difference between the first detection at the entrance and the first detection at the corridor (as represented in figure A.13 on page 156), we can interpret this as a queuing time for entering the festival area.

We now calculate the travel times, first without a filter and subsequently filtering out

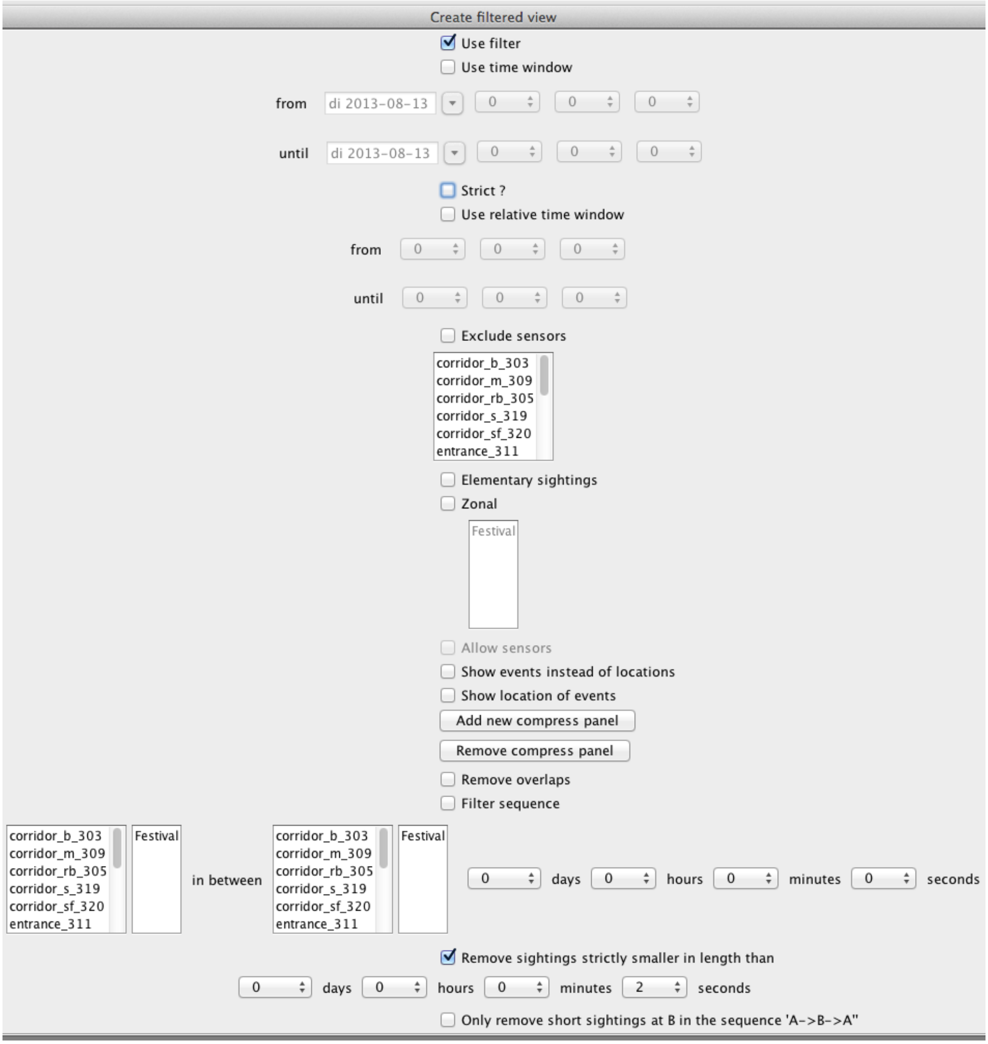


Figure A.11: The 'Create filtered view' dialog applies a *live* filter on the dataset that will be used throughout the GUI or any metadata calculations made as long as the 'Use filter' option is checked.

the solitary detections. The dialog for doing so is shown in figure A.14a on page 157. Selecting the 'Travel time between ([[)' option populates the list below with all combinations of two locations. The two opening squared brackets represent the first detections at both locations being used for the calculation. The distributions with and without filter are plotted in figure A.14a on page 157. Whereas the unfiltered distribution seems almost entirely long-tailed, filtering out the solitary detections clearly reveals a peak around 1 minute and 15 seconds. While this insight is interesting, examining one single distribution for the entire festival duration completely ignores the fact that queuing might take longer around certain

Time interval	Duration	Sightings
[16/08/2012 12:35:06-16/08/2012 12:35:07]	1 second	1 [stage_w_3...
[16/08/2012 12:36:42-16/08/2012 12:37:10]	28 seconds	1 [entrance_...
[16/08/2012 12:37:49-16/08/2012 12:38:07]	18 seconds	1 [entrance_...
[16/08/2012 12:39:15-16/08/2012 12:39:16]	1 second	1 [entrance_...
[16/08/2012 12:39:42-16/08/2012 12:39:43]	1 second	1 [entrance_...
[16/08/2012 12:40:05-16/08/2012 12:41:19]	1 minute 14 se...	1 [entrance_...
[16/08/2012 12:41:37-16/08/2012 12:42:02]	25 seconds	1 [entrance_...
[16/08/2012 12:42:21-16/08/2012 12:45:57]	3 minutes 36 s...	1 [entrance_...
[16/08/2012 12:46:15-16/08/2012 12:46:48]	33 seconds	1 [entrance_...
[16/08/2012 12:47:04-16/08/2012 12:47:23]	19 seconds	1 [entrance_...
[16/08/2012 12:48:24-16/08/2012 12:48:29]	5 seconds	1 [corridor_sf...
[16/08/2012 12:49:13-16/08/2012 12:49:14]	1 second	1 [corridor_sf...
[16/08/2012 12:49:20-16/08/2012 12:49:21]	1 second	1 [stage_mai...
[16/08/2012 12:53:30-16/08/2012 12:53:32]	2 seconds	1 [stage_mai...
[16/08/2012 12:53:49-16/08/2012 12:53:59]	10 seconds	1 [stage_mai...
[16/08/2012 12:54:56-16/08/2012 12:55:03]	7 seconds	1 [stage_mai...
[16/08/2012 12:55:26-16/08/2012 12:55:34]	8 seconds	1 [stage_mai...
[16/08/2012 12:56:05-16/08/2012 12:56:10]	5 seconds	1 [stage_mai...
[16/08/2012 12:56:50-16/08/2012 12:58:42]	1 minute 52 se...	1 [stage_mai...
[16/08/2012 13:00:17-16/08/2012 13:00:26]	9 seconds	1 [stage_mai...
[16/08/2012 13:01:17-16/08/2012 13:01:18]	1 second	1 [stage_mai...
[16/08/2012 13:03:41-16/08/2012 13:03:43]	2 seconds	1 [stage_mai...
[16/08/2012 13:04:26-16/08/2012 13:04:27]	1 second	1 [stage_mai...
[16/08/2012 13:05:34-16/08/2012 13:07:34]	2 minutes 0 sec...	1 [stage_mai...
[16/08/2012 13:08:35-16/08/2012 13:09:09]	34 seconds	1 [stage_mai...
[16/08/2012 13:09:32-16/08/2012 13:09:44]	12 seconds	1 [stage_mai...
[16/08/2012 13:10:04-16/08/2012 13:10:22]	18 seconds	1 [stage_mai...
[16/08/2012 13:10:44-16/08/2012 13:12:40]	1 minute 56 se...	1 [stage_mai...

(a)

Time interval	Duration	Sightings
[16/08/2012 12:36:42-16/08/2012 12:38:07]	1 minute 25...	1 [entrance_311]
[16/08/2012 12:40:05-16/08/2012 12:47:23]	7 minutes 18...	1 [entrance_311]
[16/08/2012 12:48:24-16/08/2012 12:48:29]	5 seconds	1 [corridor_sf_...
[16/08/2012 12:53:30-16/08/2012 12:58:42]	5 minutes 12...	1 [stage_main_...
[16/08/2012 13:00:17-16/08/2012 13:00:26]	9 seconds	1 [stage_main_...
[16/08/2012 13:03:41-16/08/2012 13:03:43]	2 seconds	1 [stage_main_...
[16/08/2012 13:05:34-16/08/2012 13:07:34]	2 minutes 0...	1 [stage_main_...
[16/08/2012 13:08:35-16/08/2012 13:13:16]	4 minutes 41...	1 [stage_main_...
[16/08/2012 13:25:19-16/08/2012 13:25:31]	12 seconds	1 [stage_main_...
[16/08/2012 13:28:42-16/08/2012 13:28:47]	5 seconds	1 [stage_main_...
[16/08/2012 13:30:08-16/08/2012 13:30:19]	11 seconds	1 [stage_main_...
[16/08/2012 13:34:17-16/08/2012 13:34:39]	22 seconds	1 [stage_main_...
[16/08/2012 13:37:21-16/08/2012 13:37:50]	29 seconds	1 [stage_main_...
[16/08/2012 13:43:35-16/08/2012 13:43:51]	16 seconds	1 [stage_main_...
[16/08/2012 13:48:39-16/08/2012 13:48:42]	3 seconds	1 [stage_main_...
[16/08/2012 13:58:16-16/08/2012 13:58:21]	5 seconds	1 [stage_cl_330]
[16/08/2012 13:59:40-16/08/2012 13:59:51]	11 seconds	1 [stage_main_...
[16/08/2012 14:07:32-16/08/2012 14:07:42]	10 seconds	1 [corridor_sf_...
[16/08/2012 14:10:39-16/08/2012 14:10:43]	4 seconds	1 [corridor_rb_...
[16/08/2012 14:14:26-16/08/2012 14:14:29]	3 seconds	1 [stage_br_317]
[16/08/2012 14:14:55-16/08/2012 14:15:42]	47 seconds	1 [stage_s_307]
[16/08/2012 14:46:08-16/08/2012 14:46:11]	3 seconds	1 [stage_br_317]
[16/08/2012 14:46:14-16/08/2012 14:47:22]	1 minute 8 s...	1 [stage_s_307]
[16/08/2012 15:00:18-16/08/2012 15:00:55]	37 seconds	1 [stage_main_...
[16/08/2012 17:17:12-16/08/2012 17:17:14]	2 seconds	1 [stage_main_...
[16/08/2012 18:43:09-16/08/2012 18:43:12]	3 seconds	1 [stage_main_...
[16/08/2012 18:53:58-16/08/2012 18:54:20]	22 seconds	1 [stage_main_...
[16/08/2012 20:34:12-16/08/2012 20:36:35]	2 minutes 23...	1 [stage_main_...

(b)

Figure A.12: Bluetooth detections without a filter (a) and with a filter removing solitary detections and compressing detection intervals within 1 minute of each other (b).

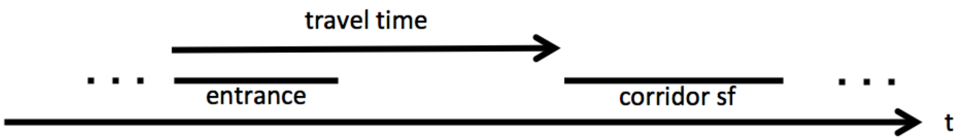
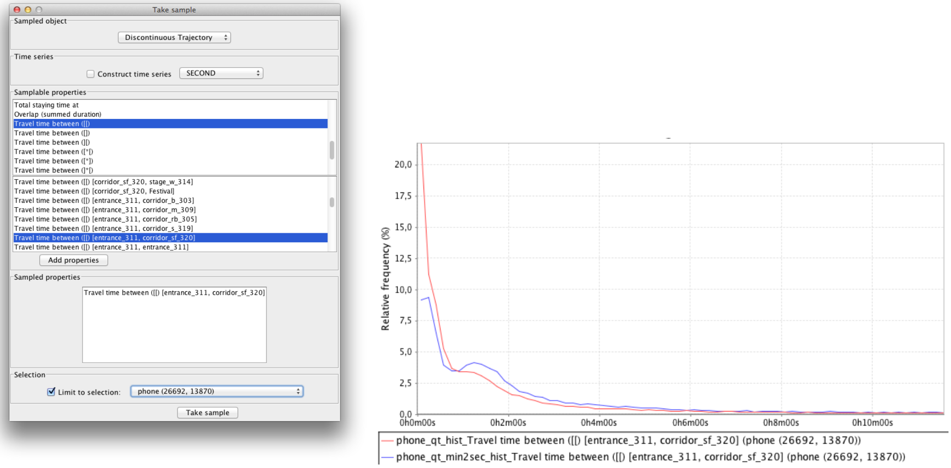


Figure A.13: Conceptual representation of the *queuing time* from the entrance to the first corridor in the festival area.

congested moments.

In order to incorporate temporal effects, we can calculate a *sample collection time series.* This is basically a sample collection where all sampled properties not only return a value for each sampled object, but also a timestamp. We can calculate a *sample collection time series* through the same dialog as for a regular sample collection, but now checking the 'Construct time series' option as shown in figure A.15 on the facing page. As soon as the calculations are finished, a new object appears under the 'Sample collection time series' branch of the 'Projects' tree. Selecting this object, the table on the right lists the sample size, minimum, maximum, average and median values for each time period. This object can also be visualized as a box-plot, which is also shown in figure A.15 on the next page. The graph shows that the queuing time usually remains below two minutes, except for the first day between 1 and 5 pm. The highest median queuing time of around 6.5 minutes is situated between 3 and 4 pm on the first day.

To investigate the movements in between the sensor locations in an aggregated way, a



(a) **(b)**

Figure A.14: Calculation of travel times between two sensor locations (a), and visualization of the histograms of the travel times distribution before (blue) and after (red) filtering (b).

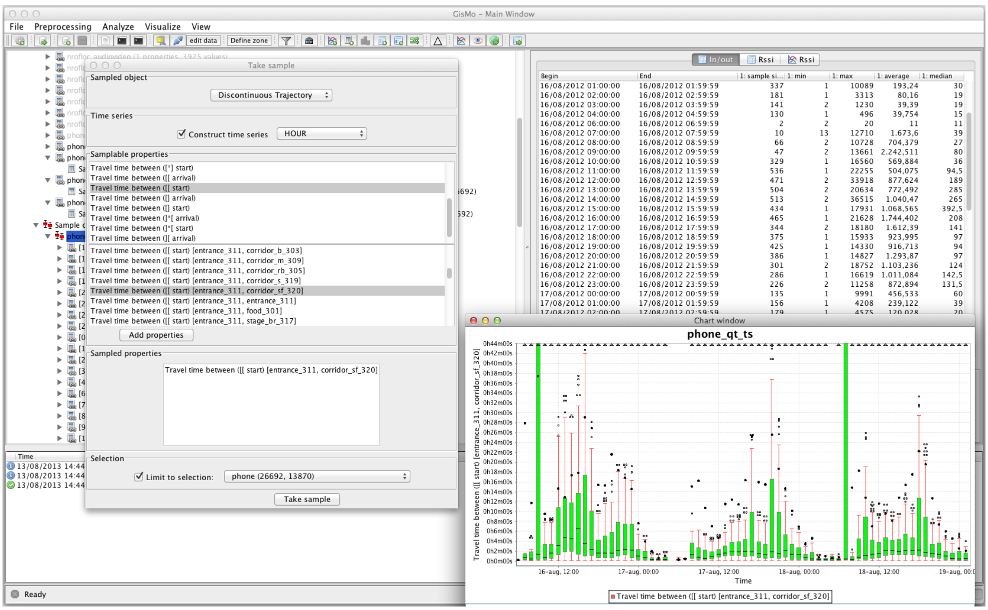


Figure A.15: Calculation of a *sample collection time series*, and its visualization through a box-plot.

flow chart can be calculated. Figure A.16 on the following page shows how certain input parameters can be set in the corresponding dialog (e.g. a minimum and maximum duration of a move, and the locations to include in the chart). The resulting flow chart is again added to the 'Projects' tree. It can be visualized as a KML (Keyhole Markup Language) file, which can be opened in Google Earth. Although not shown in the screenshot, we again used a live filter removing solitary detections. As with the sample collections, there is also an option to incorporate the temporal aspect by creating a *flow chart time series*. The flows in the resulting KML file will then automatically be associated with timestamps, allowing animations of flows to be studied.

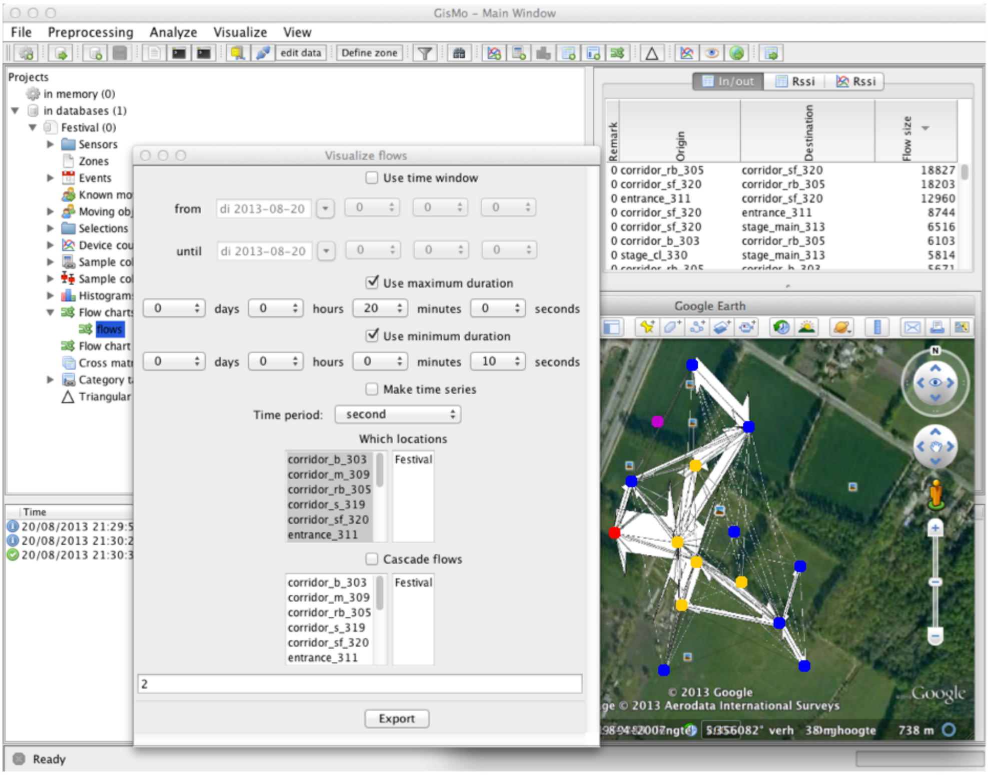


Figure A.16: Calculation of a *flow chart*, and its visualization in Google Earth as a KML file. The red circle in the flow chart represents the entrance, the blue circles the stages, the yellow circles the corridors, and the purple circle the backstage restaurant.

Besides flows, we can also visualize the Bluetooth trajectories as such. Again, the output of *GISMO* is a KML file which can be opened in Google Earth. An example is shown in figure A.17 on the next page, where two trajectories selected in the 'Projects' tree are exported to the KML file. The dialog shows several options, one of which is to color the trajectories according to the color they are assigned in *GISMO*. Colors can be assigned by double-clicking any object representing a set of one or more trajectories (i.c. trajectories, devices, selections, sensors and projects). The time slider in the Google Earth window indicates that the segments of the trajectories are also associated with a timestamp. This allows the trajectories to be viewed incrementally as they progress over time.

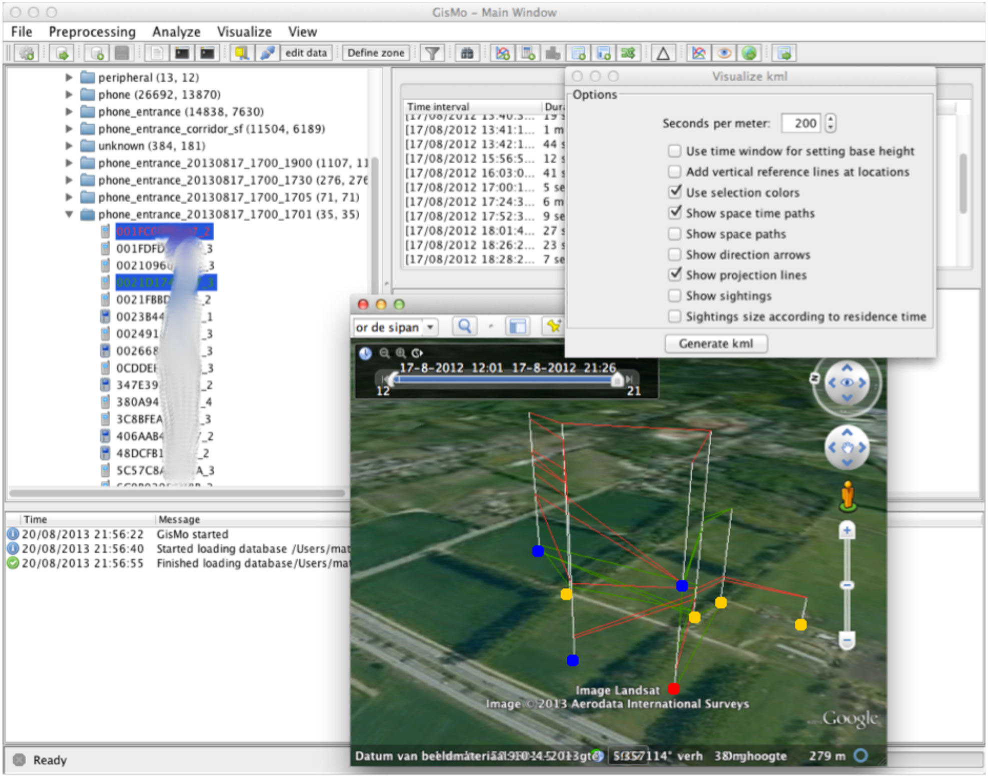


Figure A.17: Spatiotemporal visualization of two selected Bluetooth trajectories. Note that the colors in the KML file correspond to the colors in the *GISMO* user interface. The red circle in the flow chart represents the entrance, the blue circles the stages, and the yellow circles the corridors.

A.4 Final remarks

In this chapter, we presented a toolkit tailored to the analysis of episodic and proximity-based tracking data. The toolkit is named *GISMO* refering to its use as a GIS analogon for the analysis of moving objects. Although it was developed for handling Bluetooth tracking data — as demonstrated in the overview in section A.3 on page 145 — other data sources can be used as well, as long as they have the same format as described in section A.2 on page 144. For tracking data gathered through another technology than Bluetooth, one can just use dummy values of 0 for the class of device code. The user interface will then classify all devices as the type 'unknown' with no identified Bluetooth services. If the data additionally does not contain valid MAC addresses (either 17 characters long with colons, or 12 without colons), the toolkit will not be able to deduce the device brand but all other analyses will function as expected.

The toolkit's main merit is that it assembles a number of common procedures in preprocessing, selecting, transforming, analyzing and visualizing episodic proximity-based data under one accessible user interface. This way, we believe that an important contribution of the toolkit is the lowering of the learning curve to work with a data type that is becoming more common due to the growing importance of network-based tracking technologies. The live filtering option, as shown in section section A.3.2 on page 149, increases the versatility of the toolkit significantly and allows for a very thorough exploration of the data and its inherent degrees of noise and inaccuracy. The ability to export to KML files for visualization purposes increases the utility of the toolkit even further. Although not shown in section A.3 on page 145, all types of metadata can also be copied in a csv format to the terminal window the toolkit was started from. The output of *GISMO* can then be further analyzed or mined with software specifically suited for that task.

Although the toolkit has already been used for analyzing very large datasets (in the order of 100 million log lines), it does contain a potential bottleneck for even larger datasets. All data (regardless of the project being 'in memory' or 'in database'[2]) resides in RAM memory for the entire lifetime of the program. As such, the toolkit will not be able to handle datasets that are larger than the amount of internal memory of the computer it resides on. The overall performance of the toolkit is acceptable for offline analyses, but real-time environments will need a shift to a different architecture with a genuine database backend making use of indexes for increasing performance.

References

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Ahas, R., Laineste, J., Aasa, A., and Mark, U. (2007). The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In Gartner, G., Cartwright, W., and Peterson, M. P., editors, *Location Based Services and TeleCartography*, Lecture Notes in Geoinformation and Cartography, pages 445–460. Springer, Berlin.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., and Hecker, D. (2012). Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. *KI - Künstliche Intelligenz*, 26(3):241–251.

Bensky, A. (2007). *Wireless positioning technologies and applications*. Artech House, Boston, London.

[2]The directory structure associated with a database project solely serves as a backend for data permanence after program termination. All the data are still loaded in the internal memory.

Bonné, B., Barzan, A., Quax, P., and Lamotte, W. (2013). WiFiPi: Involuntary tracking of visitors at mass events. In *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Madrid.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Camagni, R., Gibelli, M. C., and Rigamonti, P. (2002). Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion. *Ecological Economics*, 40(2):199–216.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 81–88, Barcelona.

Delafontaine, M., Versichele, M., Neutens, T., and Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34:659–668.

Demšar, J., Zupan, B., Leban, G., and Curk, T. (2004). From Experimental Machine Learning to Interactive Data Mining. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D., editors, *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Computer Science*, pages 537–539. Springer, Berlin, Heidelberg.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI magazine*, 17(3):37–54.

Girardin, F. and Calabrese, F. (2008). Digital Footprinting: Uncovering Tourists with User-Generated Content. *Pervasive Computing*, 7(4):36–43.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hägerstrand, T. (1970). What about people in Regional Science? *Papers in Regional Science*, 24(1):6–21.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., and Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160:60–68.

Hainen, A. M., Remias, S. M., Bullock, D. M., and Mannering, F. L. (2013). A hazard-based analysis of airport security transit times. *Journal of Air Transport Management*, 32:32–38.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15124–9.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphic. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Jankowski, P., Andrienko, N., Andrienko, G., and Kisilevich, S. (2010). Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transactions in GIS*, 14(6):833–852.

Larsen, J. E., Sapiezynski, P., Stopczynski, A., Moerup, M., and Theodorsen, R. (2013). Crowds, Bluetooth, and Rock-n-Roll. Understanding Music Festival Participant Behavior.

Mazuelas, S., Bahillo, A., Lorenzo, R. M., Fernandez, P., Lago, F. a., Garcia, E., Blas, J., and Abril, E. J. (2009). Robust Indoor Positioning Provided by Real-Time RSSI Values in Unmodified WLAN Networks. *IEEE Journal of Selected Topics in Signal Processing*, 3(5):821–831.

Monreale, A. and Andrienko, G. (2010). Movement Data Anonymity through Generalization. *Transactions on Data Privacy*, 3(2):91–121.

Neutens, T., Versichele, M., and Schwanen, T. (2010). Arranging place and time: A GIS toolkit to assess person-based accessibility of urban opportunities. *Applied Geography*, 30(4):561–575.

Ojala, T., Hakanen, T., Mäkinen, T., and Rivinoja, V. (2005). Usage Analysis of a Large Public Wireless LAN. In *2005 International Conference on Wireless Networks, Communications and Mobile Computing*, volume 1, pages 661–667.

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A., Penn, A., Fraser, D., and Jones, T. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *8th International Conference on Ubiquitous Computing (UBICOMP 2006)*, pages 315–332, Orange County, CA.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Shaw, S.-L. and Yu, H. (2009). A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical–virtual space. *Journal of Transport Geography*, 17(2):141–149.

Shaw, S.-L., Yu, H., and Bombom, L. S. (2008). A Space-Time GIS Approach to Exploring Large Individual-based Spatiotemporal Datasets. *Transactions in GIS*, 12(4):425–441.

Shoval, N. and Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1):141–159.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., and Andrienko, N. (2011). Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In *Third ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pages 51–58, Chicago, IL. ACM.

Vanheel, F., Verhaevert, J., Laermans, E., Moerman, I., and Demeester, P. (2011). Automated linear regression tools improve RSSI WSN localization in multipath indoor environment. *EURASIP Journal on Wireless Communications and Networking*, 2011(38):1–27.

Vanheel, F., Verhaevert, J., Laermans, E., Moerman, I., and Demeester, P. (2013). Pseudo-3D RSSI-based WSN localization algorithm using linear regression. *Wireless Communications and Mobile Computing*.

Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012a). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2):208–220.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., and Van de Weghe, N. (2012b). Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10):14196–14213.

Witten, I., Frank, E., and Hall, M. A. (2010). *Data Mining: Practical machine learning tools and techniques.* 3 edition.

Yu, H. and Shaw, S. (2008). Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographical Information Science*, 22(4):409–430.

Zhou, S. and Pollard, J. (2006). Position Measurement using Bluetooth. *IEEE Transactions on Consumer Electronics*, 52(2):555–558.

B
Accomplishments

B.1 Publications (first author)

B.1.1 A1 (journal articles)

Versichele, M., Neutens, T., Delafontaine, M., Van de Weghe, N. (2012) The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, 32(2): 208-220 (published).

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., Van de Weghe, N. (2012) Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors: Tour of Flanders 2011. *Sensors*, 12(10): 14196-14213 (published).

Versichele, M., Neutens, T., Claeys Bouuaert, M., Van de Weghe, N. (2013) Time-geographic derivation of feasible co-presence opportunities from network-constrained episodic movement data. *Transactions in Geography* (in press).

Versichele, M., De Grootte, L., Neutens, T., Claeys Bouuaert, M., Moerman, I., Van de Weghe, N. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: a case study of Ghent, Belgium (under review).

Versichele, M., Neutens, T., Moerman, I., Van de Weghe, N. *GISMO*: a Geographical Information System for the analysis of Moving Objects based on episodic proximity-based sensor tracking data (under review).

B.1.2 B2 (book chapters)

Versichele, M., Neutens, T., Van de Weghe, N. (2013) Person Monitoring with Bluetooth Tracking. In *Mobility Data: Modeling, Management and Understanding*, eds. Renso, C., Spaccapietra, S., Zimányi, E., Cambridge University Press, 277-293 (in press).

B.1.3 P1 (conference proceedings)

Versichele, M., Huybrechts, R., Neutens, T., Van de Weghe, N. (2012) Intelligent Event Management with Bluetooth Sensor Networks. In *Proceedings of the 8th International Conference on Intelligent Environments (IE)*, 311-314 (published).

B.1.4 C1 (other conference papers)

Versichele, M., Delafontaine, M., Neutens, T., Van de Weghe, N. (2010) Potential and Implications of Bluetooth Proximity-Based Tracking in Moving Object Research. In *Proceedings of the first workshop on movement pattern analysis (MPA 2010)*, 111-116 (published).

B.1.5 A4 (non-academic articles)

Versichele, M., Neutens, T., Huybrechts, R., Vlassenroot, S., Gautama, S., Van de Weghe, N. (2012). Bluetooth: meer dan gadget voor mobiliteitsonderzoek - Vakgroep Geografie van UGent brengt mensenstromen in kaart. *Verkeersspecialist*, 192(december 2012), 26–29.

B.2 Publications (co-author)

Neutens, T., Versichele, M., Schwanen, T. (2010) Arranging place and time: A GIS toolkit to assess person-based accessibility of urban opportunities. *Applied Geography*, 30(4): 561-575.

Delafontaine, M., *Versichele, M.*, Neutens, T., Van de Weghe, N. (2012) Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34: 659–668 (published).

Qiang, Y., Delafontaine, M., *Versichele, M.*, De Maeyer, P., Van de Weghe, N. (2012) Interactive analysis of time intervals in a two-dimensional space. *Information Visualization*, 11(4): 255-272.

Barzan, A., Bonne, B., Quax, P., Lamotte, W., *Versichele, M.*, Van de Weghe, N. (2013). A comparative simulation of opportunistic routing protocols using realistic mobility data obtained from mass events. In *2013 IEEE 14th International Symposium on 'A World of Wireless, Mobile and Multimedia Networks (WoWMoM)'*.

B.3 Master's theses (advisor)

Table B.1: Master's theses on Bluetooth tracking.

Academic year	Student	Title
2009-2010	Bram Van Londersele	Captatie van de bewegingen van festivalgangers op Rock Werchter 2009 d.m.v. Bluetooth-tracking
2010-2011	Tim Claeys	Kan Bluetooth tracking een hulpmiddel zijn voor mobiliteitsstudies ? Case-studie : Gentse Feesten 2010
	Edward de Mûelenaere	Bluetoothtracking voor marketingdoeleinden: Shoppingcenter Gent Zuid
	Sara de Roeck	Sequentieanalyse van het bewegingspatroon van de bezoekers aan de Gentse Feesten
	Frederik Van Bossche	Wireless local positioning met Bluetooth
2011-2012	Tom Baeyens	Bluetoothtracking: Privacy en representativiteit
	Roel Huybrechts	Realtime Bluetoothtracking: Testcase Lichtfestival 2012
	Stefaan De Vos	Positioning with Bluetooth
	Stephanie Goudeseune	Mobile Bluetooth-tracking van toeschouwers op de Ronde van Vlaanderen 2011
2012-2013	Liesbeth De Groote	Sequentieanalyse van bewegingspatronen bij toeristen in Gent door middel van Bluetooth-tracking

B.4 Tracking projects

Table B.2: Tracking projects summary.

Year	Project	Number of days[a]	Number of locations[b]
2009	Rock Werchter	4	36
	Nacht van de onderzoekers	1	17
	I Love Techno	1	15
	Horeca Expo	5	20
2010	Rock Werchter	4	38
	Gentse Feesten	10	59
	Student Kick Off	1	30
	I Love Techno	1	27
	Horeca Expo	5	12
	Shopping Gent-Zuid	19	56
2011	Ronde van Vlaanderen	1	1[c]
	Gentse Feesten	10	45
	Sint-Niklaas	61	35
	Dampoort	3	8
2012	Lichtfestival	4	26
	Toerisme Gent	29	15
	Gentse Feesten	10	34
	Student Kick Off	1	12
	Pukkelpop	3	15
2013	Gentse Feesten[d]	10	38

[a] Actual duration of the running project.
[b] Both permanent and temporary locations.
[c] Mobile platform.
[d] Bluetooth and WiFi tracking.

Nederlandse samenvatting –Summary in Dutch–

Menigtes ('crowds'), en meer bepaald de bewegingen van personen die hen vormen, vormen een belangrijk onderzoeksthema in tal van domeinen. Wanneer deze menigtes als tijdruimtelijke eenheden beschouwd worden, vormen ze bijvoorbeeld een potentieel veiligheidsrisico rond flessenhalzen ('bottlenecks') wanneer een te grote dichtheid aan personen bereikt wordt. Lossere groepen van individuen met gelijkaardige intenties worden dan weer vaak bestudeerd vanuit een een micro-economische (bv. binnen een winkelcentrum) of toeristische context (bv. binnen een stadscentrum). Om deze bewegingen rechtstreeks te bestuderen of simulaties en modellen te valideren is er een grote noodzaak aan empirische gegevens. Verplaatsingen van grote groepen op kleine schalen zijn echter zeer moeilijk in kaart te brengen doordat conventionele methodes vaak een directe participatie van het bestudeerde individu inhouden en dus moeilijk te schalen zijn naar grote groepen. Indirecte observatie via camera's vormen een alternatief, maar analyses van trajecten zijn beperkt tot het zichtsveld van één camera doordat individuen onder realistische omstandigheden niet over verschillende camera's gevolgd kunnen worden.

Het meten van bewegingen van objecten die gerelateerd kunnen worden aan bewegingen van personen biedt betere uitzichten. Verschillende van deze 'proxies' zijn reeds bestudeerd, maar de laatste jaren heeft het onderzoek zich duidelijk geconcentreerd rond het gebruik van mobiele telefoons. Ondanks de vele technologische mogelijkheden, blijkt er echter nog steeds geen geschikte methodologie te bestaan om kleinschalige bewegingen van menigtes in kaart te brengen. Het anoniem traceren van verplaatsingen op basis van gegevens van mobiele operatoren biedt immers wel het voordeel dat er geen participatie nodig is van de bestudeerde individuen, maar kan slechts een ruimtelijk detailniveau aanleveren op basis van de dichtheid van het netwerk van celmasten. Vaak zijn bewegingen binnen hetzelfde captatiegebied van eenzelfde celmast net het meest relevant met betrekking tot hun geografische context. In contrast hiermee staan methodologieën op basis van GPS (global positioning system) technologie, die wel gedetailleerde informatie kan leveren maar te sterk afhankelijk zijn van (de verschillen in) de graad van participatie van de verschillende individuen.

Recent is reeds aangetoond dat andere draadloze technologieën gebruikt kunnen worden om de aanwezigheid van een mobiel toestel lokaal te detecteren. Door de grote verspreiding van de technologie over mobiele toestellen en door de vrij eenvoudige detectieprocedure, is vooral Bluetooth reeds voorgesteld als alternatief om kleinschalige bewegingen te capteren. Door het plaatsen van sensoren op strategische plaatsen binnen een studiegebied kunnen de bewegingen van detecteerbare ('discoverable') toestellen tussen deze sensoren vrij eenvoudig en op een anonieme wijze gereconstrueerd worden ('Bluetooth tracking'). Naast gemotoriseerd verkeer kunnen zo ook voetgangersbewegingen gemeten worden. Doordat de graad van ruimtelijk detail zelf te bepalen is door de plaatsing van de sensoren, en er geen participatie nodig is van het publiek lijkt de methodologie vooral uitermate geschikt voor het opmeten van verplaatsingen van grote menigtes op kleine schaal.

Ondanks deze voordelen zijn er nog maar weinig gedocumenteerde experimenten in de wetenschappelijke literatuur. Zo blijft het vooralsnog onduidelijk wat het uiteindelijke potentieel van de methodologie is. Dit proefschrift trachtte hierin een duidelijker beeld te scheppen door [i] zowel de mogelijkheden als de beperkingen van Bluetooth tracking op massa-evenementen uitgebreid te illustreren; [ii] te onderzoeken of er applicaties mogelijk zijn buiten de context van massa-evenementen; en [iii] het analyseren van Bluetooth tracking gegevens en hun specifieke eigenschappen dieper te verkennen. Vertrekkend van de wetenschappelijke literatuur werden deze objectieven vertaald naar een onderzoeksagenda bestaande uit vier onderzoeksvragen. Elk van deze onderzoeksvragen werd achtereenvolgens behandeld door een apart hoofdstuk. Twee van deze hoofdstukken zijn reeds gepubliceerd in een internationaal peer-reviewed academisch tijdschrift, één is geaccepteerd en één wordt momenteel beoordeeld. De vier onderzoeksvragen waren:

RQ 1: Welke mogelijkheden biedt Bluetooth tracking bij het bestuderen van tijdruimtelijke dynamieken binnen menigtes op massa-evenementen?

RQ 2: Kan Bluetooth technologie gebruikt worden om menigtes verpreid over een ruim gebied te tellen en in kaart te brengen?

RQ 3: Hoe kan de beweging van een menigte tussen twee sensoren in gemodelleerd worden?

RQ 4: Wat is de waarde van Bluetooth tracking buiten de context van massa-evenementen?

Hoofdstuk 2 op pagina 19 behandelde onderzoeksvraag 1 door een case study tijdens de Gentse Feesten 2010 te bespreken. Zowel het evenement als het werkingsprincipe van Bluetooth tracking werden uitvoerig besproken. Speciale aandacht ging ook naar de gebruikte Bluetooth 'scanners' en sensoren. Gedurende tien dagen werden bezoekers op 22 locaties (waarvan 11 binnen de evenementenzone) door Bluetooth-sensoren gedetecteerd, wat aanleiding gaf tot een dataset van 80.828 mobiele telefoons die 152.487 trajecten aflegden binnen de evenementenzone. Het informatiepotentieel van Bluetooth tracking voor een massa-evenement werd vervolgens geïllustreerd door verschillende analyses uit te voeren. Zo werd er via steekproeven eerst bepaald dat 11$\pm$1,8% van het publiek traceerbaar was. Met deze 'detectie-ratio' konden we ruwe schattingen maken van aantallen bezoekers. Zo werd het totale aantal bezoekers over het volledige evenement op ongeveer 1,4 miljoen geschat, wat in de lijn lag van de verwachtingen van de organisatie (1,5 miljoen). Door de gegevens van de verschillende sensoren in de feestzone te aggregeren brachten we op een gedetailleerde manier het aantal bezoekers over de tijd heen in beeld. Waar het publiek overdag verspreid is over de meeste pleinen, tekende er zich 's nachts een duidelijke concentratie af. Een volgende analyse toonde aan dat de meerderheid (65%) van de geregistreerde personen slechts één dag het evenement bezocht. Het aandeel meerdaagse bezoekers per plein varieerde onderling sterk van ongeveer 8% tot meer dan 20%. Door bezoekers ook te detecteren in de twee treinstations en de park&ride tramhalte buiten het centrum, demonstreerden we dat het aandeel bezoekers dat de trein nam vrij gelijk was over de verschillende dagen heen (5–6%), maar dat het aandeel tramgebruikers sterker varieerde per dag (3–7%). De duur van een bezoek bedroeg gemiddeld iets minder dan vier uur, maar werd gekenmerkt door een grote spreiding. Een beknopte stroomanalyse legde een karakteristiek patroon bloot waar stromen gelijkmatig verdeeld zijn overdag, maar er 's nachts duidelijke concentraties plaatsvinden gevolgd door een algemene uitstroom. Het hoofdstuk werd afgesloten met een discussie die eerst de meerwaarde van de methodologie besprak, en nadien ook focuste op problemen die nog verder onderzoek vergen. Zo werden er belangrijke vragen gesteld over de representativiteit van de verzamelde gegevens naar de volledige populatie toe (leeftijd, geslacht, onderwijs, enz.).

Waar een menigte binnen een niet te grote en duidelijk afgelijnde evenementenzone nog in kaart kan gebracht worden door een netwerk van statische Bluetooth sensoren, wordt deze benadering niet meer haalbaar wanneer een publiek een veel groter gebied bestrijkt. Er diende dan ook onderzocht te worden of en hoe Bluetooth tracking nog bruikbaar is binnen zo een scenario. Hoofdstuk 3 op pagina 45, trachtte op onderzoeksvraag 2 een antwoord te formuleren door het publiek van een wielerwedstrijd (Ronde van Vlaanderen 2011) in kaart te brengen door middel van een mobile mapping experiment waar toeschouwers langs de kant van de weg gedetecteerd werden door Bluetooth sensoren op een mobiel platform dat hetzelfde traject als de wielrenners aflegde. Een experiment vóór de wedstrijd toonde aan dat de snelheid van het mobiele platform een negatieve invloed had op het detectieproces, maar dat een klasse 1 Bluetooth sensor (dit is de meest gevoelige klasse) zelfs tegen vrij hoge snelheden geen statische toestellen langs de weg mistte. Tijdens de wedstrijd werden door twee Bluetooth sensoren bijna 16.000 telefoons gedetecteerd. Door het afgelegde traject op te delen in segmenten van 1 kilometer werd een gedetailleerd beeld van de relatieve drukte langsheen het parcours verkregen, waar drukke zones meestal meestal overeenkwamen met hellingen of kasseistroken. Door over verschillende segmenten visuele tellingen van toeschouwers (via een camera op het mobiele platform) te vergelijken met het aantal gedetecteerde Bluetooth-toestellen, werd een detectie-ratio van 14,3$\pm$3,9% met outliers, en 13,0±2,3% zonder outliers berekend. Door extrapolatie op basis van deze cijfers werd een inschatting gemaakt van het aantal toeschouwers op het drukste segment en over het hele parcours. Een verdere analyse op de dataset verzameld tijdens de wedstrijd toonde geen onmiddellijk rechtstreeks verband tussen snelheid en een aantal indicatoren van het detectieproces. Wel vertoonden beide sensoren op het platform een vrij kleine overlap in gedetecteerde toestellen, wat de impact van de exacte plaatsing van sensoren op het platform aantoont. De relatieve standaardfout van 17,9% bij het inschatten van de grootte van de menigte lag iets hoger dan de alternatieve methodologieën besproken in de literatuur, maar de grootste meerwaarde van de belichte proof-of-concept lag niet louter in het tellen van een publiek als dusdanig maar eerder in het in kaart brengen ervan, en bijkomend tijdruimtelijke inzichten te verwerven op basis van individuele trajecten.

In hoofdstuk 4 op pagina 67 werd de spatiotemporeel schaarse of episodische ('episodic') aard van Bluetooth tracking gegevens onder de loep genomen. Doordat toestellen slechts op een beperkt aantal locaties gedetecteerd worden, bestaan de tracking gegevens vaak uit lange periodes waar de locatie van een toestel niet gekend is. Er werd onderzocht of de beweging van een menigte tussen twee locaties in gemodelleerd kan worden (onderzoeksvraag 3). Hiertoe werd aansluiting gevonden met concepten uit de tijdsgeografie. Een verblijf van een toestel ter hoogte van een sensor is conceptueel immers vergelijkbaar met een vaste activiteit op een bepaalde locatie, en het modelleren van de mogelijkheden tot flexibele activiteiten tussen twee vaste activiteiten in is vergelijkbaar met het modelleren van de bewegingsvrijheid tussen twee sensoren. Op basis van het concept van een tijdruimte prisma werd een model ontwikkeld waar de gezamenlijke aanwezigheid van gedetecteerde individuen tussen twee sensoren over een netwerk heen berekend kan worden. Door de afwijking ten opzichte van het kortste pad te beperken werden van deze ontmoetingsmogelijkheden enkel de meest realistische ('feasible') overgehouden. Het model werd toegepast op de gedetecteerde bezoekers van het Lichfestival te Gent in 2012. Door de ontmoetingsmogelijkheden van alle bezoekers op verschillende tijdstippen te berekenen, werd de beweging van de menigte langsheen het uitgestippelde bezoekersparcours gereconstrueerd. De output van het model kon als een inschatting van de potentiële drukte geïnterpreteerd worden, maar verdere gegevens dienden de nauwkeurigheid van de bekomen simulaties te controleren.

Waar onderzoeksvragen 1–3 uitgaan van het gebruik van Bluetooth tracking voor het bepalen van posities van tijruimtelijk verenigde menigtes, verlegt onderzoeksvraag 4 de focus naar groepen personen met vergelijkbare intenties. In hoofdstuk 5 op pagina 89 werd de probleemstelling concreet naar een toeristische context vertaald. Gedurende twee weken werden bezoeken aan de voornaamste toeristische attracties in Gent geregistreerd door Bluetooth sensoren. Bijkomend werd een deel van de bezoekers als hotelgast geïdentificeerd door sensoren in een aantal hotels te plaatsen. Er werd aangetoond dat activiteiten slechts uit detectiegegevens afgeleid kunnen worden nadat het aanwezige ruis in de sensorgegevens weggefilterd werd. Nadien werden eerst verschillende bezoekerssegmenten afgeleid op basis van onder andere het verschil tussen bezoeken aan open en gesloten attracties, en het al dan niet geïdentificeerd zijn als hotelgast. Voor elk van deze bezoekerssegmenten werd dan gekeken naar welke attracties vaker gecombineerd bezocht werden door middel van een associatieregel-analyse. De resultaten werden gecombineerd geografisch in kaart gebracht door middel van 'bezoekpatroonkaarten' ('visit pattern maps'). Ondanks de noodzaak aan filters en de beperkte tijdsduur van het tracking experiment, werd aangetoond dat door de combinatie van Bluetooth tracking en een data-mining techniek zoals associatieregel-analyse interessante en waardevolle inzichten verworven kunnen worden. In de discussie werd verder besproken hoe de resultaten op zowel korte als langere termijn gebruikt kunnen worden voor een beter toeristisch beheer.

In appendix A op pagina 141 werd de *GISMO*-toolkit voorgesteld. Deze software werd gedurende de laatste vier jaar ontwikkeld om de verzamelde Bluetooth tracking gegevens te verwerken. De verschillende mogelijkheden van de toolkit werden uitvoerig geïllustreerd door een dataset verzameld op een muziekfestival te verwerken.

In de afsluitende discussie in chapter 6 op pagina 121 werden eerst de wetenschappelijke bijdrages van de verschillende studies opgelijst. Daarna werd er dieper ingegaan op enkele aspecten van de Bluetooth tracking methodologie die nog verder onderzocht dienen te worden. Zo zal een diepgaandere interpretatie van bewegingen van menigtes extra informatie over getrackte individuen noodzaken, maar de exacte implementatie van een combinatie van Bluetooth tracking met kwalitatieve methodes zoals interviews blijft onduidelijk. Verder blijkt de Bluetooth detectie-ratio over de jaren heen gezakt te zijn, en dient de toekomst uit te wijzen of Bluetooth-technologie vervangen of gecombineerd moet worden door/met andere draadloze technologieën zoals bijvoorbeeld WiFi. Een enquête in 2013 leverde de eerste aanwijzingen voor een zelfselectie bias ('self selection bias') op basis van geslacht en leeftijd, maar verder onderzoek is nodig om de tijdruimtelijke variabiliteit van deze bias te doorgronden. Ten slotte werd ingegaan op privacy-aspecten van de methodologie, zowel op legaal als ethisch vlak.

Curriculum vitae

Mathias Versichele (°1984, Lokeren) finished high school at the Pius X-Institute in Zele in 2001. Subsequently, he received a Master's degree in 'bioscience engineering: environmental technology' in 2007 (*magna cum laude*) and a Master's degree in 'geography and geomatics' in 2009 (*summa cum laude*) at Ghent University. In 2010, he received a grant from the Agency for Innovation by Science and Technology (IWT) and became a PhD-fellow at the Department of Geography at Ghent University. His research on 'Bluetooth tracking' resulted in several participations in major international conferences and the publication of several articles in international academic journals. In 2012, he received an award for the Best Poster Award at the International Conference on Intelligent Environments (IE2012).